SOME CONTRIBUTIONS IN COMPLEX STATISTICAL DATA ANALYSIS FOR CLASSIFICATION, FEATURE SELECTION, AND HUMAN FERTILITY MODELS

David A. Pritchard

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biostatistics.

Chapel Hill
2020

Approved by:

Yufeng Liu

Matt Psioda

Quefeng Li

Naim Rashid

Anne Steiner

ABSTRACT

David A. Pritchard: Some Contributions in Complex Statistical Data Analysis
(Under the direction of Yufeng Liu and Matthew Psioda)

The overall goal of this dissertation is twofold: (i) to develop methodologies in two related fields of machine learning, namely classification and variable selection, and (ii) to improve upon the available statistical methods for the study of human fertility.

For the area of machine learning, we first present a new classification method, *composite quantile-based classifiers*. It was established in Hennig and Viroli (2016) that for univariate data and under some assumptions, the decision rule based upon the distances of an observation to the corresponding within-class quantiles for the optimal choice of quantile levels is the Bayes rule. Conceptually, our goal is to use these most powerful univariate classifiers as building blocks for a multivariate classifier. In brief, we propose aggregating the component-wise distances from the feature vector of the observation to the within-class quantiles corresponding to the one-at-a-time optimal choices of the quantile level. Aggregation is performed through an appropriately chosen linear combination of the component-wise distances. Our second contribution in machine learning is in the development of filtering and variable selection methods in the field of bioactive compound identification. Historically, compounds found in plant tissues have been fruitful bases for deriving compounds that are deleterious to certain pathogens and cancer cell lines. Our research constructs a pipeline for processing raw data obtained through liquid chromatography - mass spectrometry (LC-MS) technology and obtaining a ranking of potential bioactive compounds.

For the area of human fertility studies, we consider data in the form of day-specific covariates, with pregnancy as the outcome variable. We present research aimed at overcoming several limitations in the literature, including the ability to account for continuous covariates and to allow for missing data. Furthermore, we incorporate the assumed true unimodal shape of the fertile window day probabilities into the model with the goal of reducing the variance of the posterior distribution.

TABLE OF CONTENTS

LIST OF TABLES x

LIST OF FIGURES xii

CHAPTER 1: INTRODUCTION AND LITERATURE REVIEW 1

1.1 Introduction 1

1.2 Machine learning topics 1

1.2.1 Supervised learning: classification 2

1.2.1.1 Quantile-based classifiers 3

1.2.1.2 The FANS classifier 4

1.2.1.3 Novel research: composite quantile-based classifiers 5

1.2.2 Variable selection and subset identification 5

1.2.2.1 Stepwise selection methods 5

1.2.2.2 Shrinkage methods 6

1.2.2.3 Novel research: anti-pathogen compound discovery 8

1.3 Fertility models 9

1.3.1 Cycle-level time scale 10

1.3.1.1 Discrete-time proportional hazards model 10

1.3.2 Day-level time scale 10

1.3.2.1 Novel research: extending the day-specific probabilities model 12

CHAPTER 2: COMPOSITE QUANTILE-BASED CLASSIFIERS 13

2.1 Quantile classifiers for univariate data 13

2.1.1 The quantile classifier 13

2.1.2 The multimodal quantile classifier 15

2.1.3 Optimality and consistency of the multimodal quantile classifier 17

2.2 Quantile-based classifiers for multivariate data 18
2.2.1 Composite quantile-based classifiers 19
2.2.1.1 Properties of composite quantile-based classifiers 20
2.2.1.2 Augmented composite quantile-based classifiers 22
2.2.1.3 Multimodal composite quantile-based classifiers 23
2.2.2 Classifier model selection 24
2.2.2.1 Choice of quantile levels 24
2.2.2.2 Choice of linear combination coefficients 25
2.2.2.3 Connections to existing methods 25
2.2.3 Algorithm for composite quantile-based classifiers 28
2.2.4 Alternative model selection algorithm: simultaneous selection 29
2.2.5 Classifier examples 30
2.2.5.1 Example 1: Two dimensional Gaussian data 30
2.2.5.2 Example 2: Two dimensional beta and gamma distributed data 31
2.3 Numerical results 32
2.3.1 Simulation examples 33
2.3.2 Application to spam email classification 35
2.4 Discussion 38
CHAPTER 3: BIOACTIVE PEPTIDE DISCOVERY 40
3.1 Data generation 42
3.1.1 Explanatory data generation 42
3.1.2 Outcome data generation 43
3.2 Processing and filtering the explanetory data 44
3.2.1 Recovering the candidate compounds 44
3.2.2 Filtering candidate compounds 46
3.2.2.1 Selecting the fraction region of interest 46
3.2.2.2 Filtering criteria selection 47
3.3 Candidate compound ranking and validation 49

3.3.1 Candidate compound ranking procedure 49

3.3.2 Pipeline validation using cyO2 49

3.4 Discussion 50

CHAPTER 4: DAY-SPECIFIC PROBABILITIES OF CONCEPTION 52

4.1 Introduction 52

4.1.1 The biology of human fertilization 52

4.1.1.1 The fertile window 53

4.1.1.2 Data collection 53

4.1.2 Time to Conceive Study 54

4.1.3 Upcoming sections 54

4.2 The day-specific probabilities model 55

4.2.1 Model specification 55

4.2.1.1 Auxiliary model specification 56

4.3 Estimating fertile window day regression terms 57

4.3.1 Prior specification of the fertile window regression coefficients 57

4.3.1.1 Specification of the fertile window decay function S 58

4.3.1.2 Specification of the peak intensity parameter μ 59

4.3.1.3 Specification of the coefficient variance parameter ν 60

4.3.1.4 Specification of the peak intensity day parameter m 60

4.3.2 Updating step for the fertile window day coefficients and prior parameters 60

4.3.2.1 Proposal distributions 60

4.4 Missing data imputation 61

4.4.1 Gibbs step for missing X_{ijk} 61

4.4.2 Gibbs step for missing covariates 62

4.4.2.1 Gibbs step for missing continuous covariate 63

4.4.2.2 Gibbs step for missing categorical covariate 64

4.5 Simulation studies 64

4.5.1 Comparing the peak-intensity model 64

4.5.1.1 Correctly specified mean prior distribution setting 65

4.5.1.2 Shifted and scaled mean prior distribution setting 67

4.5.2 Imputing missing intercourse status 68

4.5.2.1 Missing intercourse status fertile window model only 68

4.5.2.2 Missing intercourse status fertile window plus covariates model 71

4.6 Real Data Analysis 72

4.7 Conclusion 77

APPENDIX A: COMPOSITE QUANTILE-BASED CLASSIFIERS TECHNICAL DETAILS 78

A.1 Quantile classifier further discussion 78

A.1.1 Example differences of quantile distances 78

A.1.2 Consistency of the optimal quantile classifier 78

A.1.3 Calculating the quantile classifier for a fixed quantile level 79

A.1.4 Calculating the empirically optimal quantile classifier 81

A.1.4.1 An example calculation 81

A.1.4.2 Quantile classifier algorithm 83

A.1.5 Calculating composite quantile-based classifiers 86

A.2 Theoretical proofs 87

A.2.1 Proofs of Supplementary Propositions 1-6 87

A.2.2 Proof of Theorems 1, 2, and 4 94

A.2.3 Proof of Theorem 3 98

A.2.3.1 Preliminary propositions 99

A.2.3.2 Proof of Theorem 3 102

A.2.3.3 Proofs of preliminary propositions 104

A.3 Extended simulation results 109

A.3.1 Simulated data scenarios 109

A.3.2 Classifier implementations and settings 110

A.3.3 Simulation results 111

A.3.4 Categorical and mixed data 115

APPENDIX B: DAY-SPECIFIC PROBABILITIES OF CONCEPTION TECHNICAL DETAILS 127

B.1 Complete derivations from the main paper 127

B.1.1 Updating step for the fertile window day coefficients and prior parameters 127

B.1.1.1 Updating step for γ_k 127

B.1.2 Prior specification of the fertile window peak intensity coefficient 128

B.1.2.1 Updating step for μ 129

B.1.3 Prior specification of the variance parameter ν 130

B.1.3.1 Updating step for ν 130

B.1.4 Prior specification of the index of the peak intensity day 131

B.1.4.1 Updating the peak intensity day index 131

B.1.5 Gibbs step for missing covariates 132

B.1.5.1 Gibbs step for missing continuous covariate 132

B.1.5.2 Gibbs step for missing categorical covariate 134

B.2 Detailed simulation studies 135

B.2.1 Simulation setup and model comparisons 135

B.2.2 Correctly specified prior parameters setting 136

B.2.3 Scaled and shifted prior parameters setting 138

B.2.4 Incorrect decay prior parameters setting 139

B.3 The day-specific probabilities model 158

B.3.1 Model specification 158

B.3.1.1 Computation consideration 160

B.3.2 Marginal probability of conception 160

B.3.2.1 Day-specific marginal probability of conception 162

B.3.3 Prior specification 162

B.3.3.1 Monotone effects 163

B.3.4 Posterior computation 164

B.3.4.1 Verifying the equivalence of the data augmentation model 164

B.3.4.2 The full likelihood 164

B.3.4.3 The full conditional distributions 165

B.3.5 Symmetric distributions verfication 170

B.3.5.1 Computational considerations 172

B.4 Extending the algorithm to continuous predictors 174

BIBLIOGRAPHY 178

LIST OF TABLES

2.1 Simulation study: misclassification results for block transformed data. 36

2.2 Simulation study: misclassification results for varying within-class distributional shapes 37

2.3 Misclassification rates for the spam email data set for varying numbers of observations used as training data. 38

3.1 Cycloviolacin O2 rankings 50

4.1 Correctly specified prior parameters, 5 fertile window days, $n = 200$ 69

4.2 Scaled and shifted prior parameters, 7 fertile window days, $n = 300$ 70

4.3 Imputing sexual intercourse status without covariates, 7 fertile window days, correlated missingness mechanism, $n = 300$ 73

4.4 Imputing sexual intercourse status with covariates, 5 fertile window days, independent missingness mechanism, $n = 300$ 74

4.5 Analysis of the Time to Conceive data 76

A.1 Gaussian data, correlation coefficient = 0.8 117

A.2 Exponentiated Gaussian data, correlation coefficient = 0.8 118

A.3 Block transformed data, correlation coefficient = 0 119

A.4 Block transformed data, correlation coefficient = 0.8 120

A.5 Beta distributed data 121

A.6 Gamma distributed data 122

A.7 Gaussian data with partially equivalent marginals, correlation coefficient = 0.8 123

A.8 Mixture Gaussian data, correlation coefficient = 0 124

A.9 Model fitting comparison: Gaussian data, correlation coefficient = 0 125

A.10 Model fitting comparison: Exponentiated Gaussian data, correlation coefficient = 0 126

B.1 Correctly specified prior parameters, 5 fertile window days, $n = 100$ 140

B.2 Correctly specified prior parameters, 5 fertile window days, $n = 200$ 141

B.3 Correctly specified prior parameters, 5 fertile window days, $n = 300$ 142

B.4 Scaled and shifted prior parameters, 5 fertile window days, $n = 100$ 143

B.5 Scaled and shifted prior parameters, 5 fertile window days, $n = 200$ 144

B.6 Scaled and shifted prior parameters, 5 fertile window days, $n = 300$ 145

B.7 Incorrect decay prior parameters, 5 fertile window days, $n = 100$ 146

B.8 Incorrect decay prior parameters, 5 fertile window days, $n = 200$ 147

B.9 Incorrect decay prior parameters, 5 fertile window days, $n = 300$ 148

B.10 Correctly specified prior parameters, 7 fertile window days, $n = 100$ 149

B.11 Correctly specified prior parameters, 7 fertile window days, $n = 200$ 150

B.12 Correctly specified prior parameters, 7 fertile window days, $n = 300$ 151

B.13 Scaled and shifted prior parameters, 7 fertile window days, $n = 100$ 152

B.14 Scaled and shifted prior parameters, 7 fertile window days, $n = 200$ 153

B.15 Scaled and shifted prior parameters, 7 fertile window days, $n = 300$ 154

B.16 Incorrect decay prior parameters, 7 fertile window days, $n = 100$ 155

B.17 Incorrect decay prior parameters, 7 fertile window days, $n = 200$ 156

B.18 Incorrect decay prior parameters, 7 fertile window days, $n = 300$ 157

LIST OF FIGURES

2.1 Example multimodal quantile classifier partitioning 17

2.2 Composite quantile-based classifier for Gaussian distributions setting. 31

2.3 Composite quantile-based classifier for beta- and gamma-distributed data settings. 32

3.1 Liquid Chromotography Representation 43

3.2 *Viola odorata* peptide library bioactivity 48

4.1 The elicited decay curve 59

A.1 Example within-class quantile distances and differences 79

A.2 Example empirical classification rate 82

A.3 Example partitioning of the domain of Π_0 and Π_1 99

A.4 Example mixture distribution with both categorical and continuous values 115

CHAPTER 1: INTRODUCTION AND LITERATURE REVIEW

1.1 Introduction

The overall goal of this research is twofold: (i) to develop methodologies in two fields of machine learning, namely classification and data mining, and (ii) to improve upon the available statistical methods for the study of human fertility. Since the subject matter in this dissertation takes place in two rather distinct domains, we present the introduction and literature review in two separate sections accordingly. In Section 1.2 we introduce two subfields of machine learning related to our research and present our proposed contributions, and in Section 1.3 we introduce the field of statistical fertility models and present our proposed contributions.

1.2 Machine learning topics

Machine learning is an important area of research for data analysis. In this dissertation, data are provided in the form of inputs or features, and in some settings also include outcome information. One important goal of machine learning research is to provide methodology to answer problems such as predicting the outcome values for new data with a particular set of inputs, identifying important features in the data, and observing clusters of data with similar features. There are two main types of machine learning: supervised and unsupervised learning. Supervised learning includes problems where both the inputs and the outcome variables are available, while unsupervised learning is the type of learning that occurs when no outcome variable is available. Other important branches of machine learning that don't fit neatly into these categories include semi-supervised learning, in which some but not all of the outcome data are available, and reinforcement learning, in which learning occurs through interactions with a particular environment.

In two of the chapters in this dissertation we propose machine learning methods. In Chapter 2, we propose a classification method named *composite quantile-based classifiers*. Classification falls under the more general category of supervised learning. In Section 1.2.1 we present a summary of the existing literature in the field of classification. Two existing methods that are related to composite quantile-based classifiers are presented in more detail, namely the methods proposed by Hennig and Viroli (2016) and Fan et al. (2016). An overview of composite quantile-based classifiers itself is also presented.

1.2.1 Supervised learning: classification

A classifier solves a common type of supervised learning problem where the goal is to build a rule for predicting the class membership of an observation based on a set of features. The rule is constructed from a training dataset consisting of the class membership and features for each training sample. There are numerous applications for classification, including disease classification, image or sound recognition, object discrimination, and spam detection, among many others.

Classification has a long and historied place in the literature. Some important early methods include naive Bayes, linear and quadratic discriminant analysis, nearest neighbor methods (Cover and Hart, 1967), and logistic regression. Examples of more recent methods include kernel smoothing (Mika et al., 1999), neural networks (Ripley, 1994), and mixture discriminant analysis (Hastie and Tibshirani, 1996). These methods often perform well in the classical low dimensional setting, where the number of features is smaller than the training data sample size. However, in high dimensional settings some of these methods can have identifiability issues, be computationally demanding, or suffer from poor performance due to the curse of dimensionality.

Various methods have been proposed for classification in the high-dimensional setting. The support vector machine (SVM) (Cortes and Vapnik, 1995) and penalized logistic regression (Park and Hastie, 2007) are two well-known approaches that have been successfully applied in many such settings. More generally, methods designed for the high dimensional setting often employ techniques such as shrinkage, dimension reduction, or distance-based methods, sometimes further restricting attention to the marginal information provided by the features in the data. Techniques performing shrinkage include penalized logistic regression and penalized linear discriminant analysis (Tibshirani et al., 2002; Clemmensen et al., 2011; Witten and Tibshirani, 2011). Dimension reduction is another valuable technique that is employed by methods such as sure independence screening (Fan and Lv, 2008) and supervised principal components analysis (Bair et al., 2006). Nonparametric distance-based classifiers such as centroid-based classifiers (Tibshirani et al., 2002) and median based classifiers (Jörnsten, 2004; Ghosh and Chaudhuri, 2005) have been proposed, as well as component-wise median-based classifiers (Hall et al., 2012) and quantile-based classifiers (Hennig and Viroli, 2016).

In Chapter 2 we continue to explore quantile-based classifiers as introduced in Hennig and Viroli (2016) in the high-dimensional setting. This family of classifiers is based upon a comparison of the component-wise distances of the feature vector of an observation to the within-class quantiles. The quantile-based family of methods then classifies an observation as belonging to the class which has the smaller aggregated component-wise distance from the feature vector of the observation to the within-class quantiles, where the aggregated distance is defined as the sum of the individual distances. The quantile-based classifiers in Hennig and Viroli (2016) are a single-parameter family of classifiers with the parameter specifying a common quantile level for each component at which to compare the component-wise distances of an observation to. This classification approach can work well in certain settings, in particular for the setting where the optimal choice of the quantile level is the same for each component. However, this restriction on the choice of quantile level can lead to a loss of efficiency in settings where the optimal choice of the quantile level varies across components. This naturally leads to the question of whether there is a way to allow for more flexibility in the choice of quantile levels in order to achieve improved performance in these other scenarios.

It was established in Hennig and Viroli (2016) that for univariate data and under some assumptions, the decision rule based upon the distances of an observation to the corresponding within-class quantiles for the optimal choice of quantile levels is the Bayes rule. This result motivates the methodologies developed in the research presented in Chapter 2. Conceptually, our goal is to use these most powerful univariate classifiers as building blocks for a multivariate classifier. In brief, we propose aggregating the component-wise distances from the feature vector of the observation to the within-class quantiles corresponding to the one-at-a-time optimal choices of the quantile level. Aggregation is performed through an appropriately chosen linear combination of the component-wise distances.

In the following section we describe the *quantile-based classifiers* method proposed by Hennig and Viroli (2016) in some more detail. We also describe another related classification method to called *FANS* that was proposed in Fan et al. (2016).

1.2.1.1 Quantile-based classifiers

Consider the check loss (also known as the quantile loss) function defined as

$$\rho_\theta(u) = \mathbb{1}\left(u > 0\right)\theta\, u + \mathbb{1}\left(u \leq 0\right)(1-\theta)\,(-u) = u\left[\theta - \mathbb{1}\left(u \leq 0\right)\right],$$

for some choice of quantile level $\theta \in (0,1)$. Next, consider two populations Π_0 and Π_1 with corresponding distribution functions F_0 and F_1 each on $\mathbb{R}^p$, and further let F_{0j} and F_{1j} denote the marginal distribution functions for the j-th variable with respect to F_0 and F_1, $j = 1, \ldots, p$. Then we denote the quantile distance of a point $z \in \mathbb{R}$ to the θ-th quantile of a population's j-th component as

$$\Phi_{ij}(z,\theta) = \rho_\theta\Big(z - F_{ij}^{-1}(\theta)\Big), \qquad i = 0, 1, \quad j = 1, \ldots, p.$$

Then the difference of the quantile distances of a point z to the populations' θ-th quantile of the j-th component is defined as

$$\Lambda_j(z,\theta) = \Phi_{1j}(z,\theta) - \Phi_{0j}(z,\theta), \qquad j = 1, \ldots, p.$$

The classification method proposed by Hennig and Viroli (2016) then is defined as the decision rule based upon

$$\sum_{j=1}^{p} \Lambda_j(z_j, \theta) > 0.$$

The choice of quantile level θ is chosen based upon the best empirical performance for a training sample.

1.2.1.2 The FANS classifier

Consider two classes with corresponding class conditional densities f_0 and f_1. Then the FANS classifier proposed in Fan et al. (2016) is defined by the decision rule based upon

$$\mathcal{D}_{\text{FANS}} = \left\{ \boldsymbol{z} : \ \alpha_0 + \alpha_1 \log \frac{f_{01}(z_1)}{f_{11}(z_1)} + \cdots + \alpha_p \log \frac{f_{0p}(z_p)}{f_{1p}(z_p)} \ = \ 0 \right\}.$$

The classifier then estimates class conditional marginal densities f_{ij} using nonparametric kernel density estimation, and selects the choice of coefficients $\alpha_0, \ldots, \alpha_p$ by using the penalized logistic regression coefficient estimates obtained from the transformed training data where the transformation is taken to be $x_{ij}^* = \log \frac{\hat{f}_{0j}(x_{ij})}{\hat{f}_{1j}(x_{ij})}$ for all i, j.

1.2.1.3 Novel research: composite quantile-based classifiers

In Chapter 2, we present a binary classification method that we call *composite quantile-based classifiers*. In Section 2.1, we first review the quantile classifier, a univariate distance-based classification method and its sample version. Properties of the quantile classifier are discussed, and consistency of the empirically optimal quantile classifier is established. An algorithm with which to calculate the estimated classification rate for the family of quantile classifiers as a function of the quantile level is presented. In Section 2.2, the *composite quantile-based* family of classification methods is presented. Connections to other related classifiers are discussed. Consistency of the empirically optimal composite quantile-based classifier is established. A new method for selection of a composite quantile-based classifier based on training data is proposed, and a corresponding algorithm is presented. Properties and the form of composite quantile-based classifiers are discussed. In Section 2.3, the competitive performance of these approaches is demonstrated using simulation studies as well as on a benchmark email spam application. Proofs of theoretical results are presented in the Appendix.

1.2.2 Variable selection and subset identification

Variable selection is a common problem in statistics. There are several reasons for which variable selection is desired, including prediction accuracy, interpretation, and variable identification.

1.2.2.1 Stepwise selection methods

One common technique for variable selection is best-subset selection. This involves considering all possible combinations of variables and choosing the combination that offers the best score by some metric. Two commonly used metrics are the Akaike Information Criterion (AIC) (Akaike, 1974) and the Bayesian information criterion (BIC) (Schwarz et al., 1978). The AIC is defined as

$$2k - 2\log(\hat{L}),$$

where $\hat{L}$ is the observed value for the maximum likelihood of the model, and k is the number of parameters in the model. The BIC is defined as

$$k\log(n) - 2\log(\hat{L}),$$

where n is the number of observations in the data, and $\hat{L}$ and k are as before. In both cases, the criteria penalize the number of parameters in the model. Although best-subset selection is an appealing strategy, the number of subsets that need to be considered is exponential in k, and is typically not considered computationally feasible for more than 40 parameters Hastie et al. (2009).

Rather than search through all possible subsets of variables as in best-subset selection, another approach is to consider a greedy variable selection algorithm the chooses or discards variables one-at-a-time. Two such examples of this technique are forward-stepwise selection and backward-stepwise selection. Forward-stepwise selection takes the approach of one-at-a-time adding in the variable from the remaining set of variables that are not yet in the model, and where the variables chosen is the one that best improves the model by some metric such as the observed maximum likelihood value. Backwards selection works similarly, except that you start with all of the variables in the model, and then remove variables one-at-a-time by removing the variable that results in the best result with respect to some metric such as the observed maximum likelihood value. It is clear that forward- or backward-stepwise selection will result in models that fit the observed data no better than the best-subsets model with the same number of variables, however they are much less computationally expensive to obtain.

A hypothesis testing-based approach to variable selection is proposed in (Gong et al., 2018). This approach tests against the null hypothesis at each step that the response is uncorrelated with the remaining variables given a set of selected variables. The approach is flexible in that any variable selection procedure can be used and provides the ability to determine a stopping point for the procedure.

1.2.2.2 Shrinkage methods

While subset selection results in models with a reduced number of variables, the process is discrete in the sense that a variable is either in or out of a model. Another approach is to penalize the magnitude of the model coefficient values, which has the effect of shrinking their values towards zero. Furthermore, certain choices of penalization parameters have the effect of causing some of the parameters to have a value of exactly zero, resulting in another form of variable selection.

One such penalization approach is called lasso penalization (Tibshirani, 1996). This approach is defined for linear regression as

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} ||(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})||^2 \text{ subject to } \sum_{j=1}^{p} |\beta_j| \leq t,$$

where t is a nonnegative tuning parameter. An equivalent representation of the lasso estimator is given by

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left\{ \frac{1}{2} \sum_{i=1}^{n} ||(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})||^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\},$$

and where λ is a nonnegative tuning parameter. Thus, as the value of λ increases, it results in a continuous form of variable selection. A variant of the lasso penalty is the elastic net penalty, which is defined as

$$\lambda \sum_{j=1}^{p} \left(\alpha \beta_j^2 + (1 - \alpha) |\beta_j| \right),$$

and where α is a turning parameter taking a value between zero and one, inclusive. The elastic net has the same behavior of shrinking variables to exactly zero, but also tends to keep groups of correlated variables in the model together.

Another related method to the lasso is least angle regression (LAR) (Efron et al., 2004). The LAR method is defined in terms of the following algorithm.

1. Standardize the predictors to have mean zero and norm one, and start with the residuals $\boldsymbol{r} = \boldsymbol{y} - \bar{\boldsymbol{y}}$ and $\beta_1 = \cdots = \beta_p = 0$.

2. Find the predictor $\boldsymbol{x}_j$ most correlated with $\boldsymbol{r}$.

3. Move β_j from zero towards its least-squares coefficient for $\boldsymbol{x}_j$ regressed on $\boldsymbol{r}$, until some other competitor $\boldsymbol{x}_j$ has as much correlation with the current residual as does $\boldsymbol{x}_j$.

4. Let $\mathcal{A}_k$ be the set of nonzero coefficients and $\boldsymbol{r}_k$ the residual at step k, and define $\boldsymbol{\delta}_k = (\boldsymbol{X}_{\mathcal{A}_k}^T \boldsymbol{X}_{\mathcal{A}_k})^{-1} \boldsymbol{X}_{\mathcal{A}_k} \boldsymbol{r}_k$. Then for the $(k+1)$-th step, move all of the coefficients corresponding to the set $\mathcal{A}_k$ in the direction of $\boldsymbol{\delta}_k$ until some other variable has as much correlation with the current residual. Repeat this step until completion.

The LAR algorithm has the following connection to lasso. If, whenever a nonzero coefficient becomes zero, then it is dropped from the active set of variables and the direction $\boldsymbol{\delta}_k$ is subsequently recomputed, then LAR provides the full lasso path.

Another method that is similar in spirit to the lasso albeit with a somewhat different formulation is the Dantzig selector (Candes et al., 2007). The method is defined as the solution to

$$\arg\min_{\boldsymbol{\beta}} ||\boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})||_\infty \text{ subject to } ||\boldsymbol{\beta}||_1 \leq t,$$

for some nonnegative tuning parameter t. Thus the Dantzig selector replaces squared error loss by the maximum absolute value of its gradient. The solution to this expression can be shown to be a linear programming problem.

In some problems the predictor variables belong to a predefined group such as a collection of dummy-coded variables used to represent the levels of a categorical variable, or a set of variables that represent a collection of related biological measurements. In that case it is usually desirable to shrink and select the members of such a group together. One method that provides this behavior is the grouped lasso (Bakin et al., 1999; Lin et al., 2006). Suppose that p predictor variables are divided into L groups with p_ℓ variables in the ℓ-th group. Further define $\boldsymbol{X}_\ell$ and $\boldsymbol{\beta}_\ell$ to be the design matrix and coefficient vector corresponding to the ℓ-th group. Then the grouped lasso is defined as

$$\arg\min_{\boldsymbol{\beta}} \left\{ ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||^2 + \lambda \sum_{\ell=1}^{L} \sqrt{p_\ell} ||\boldsymbol{\beta}_\ell|| \right\}.$$

Thus sparsity is encouraged at both a group and an individual level since individual members are penalized for being members of a group by the $\sqrt{p_\ell}$ term.

1.2.2.3 Novel research: anti-pathogen compound discovery

One application of variable selection is in the discovery of explanatory variables with an effect on the outcome of interest. As a collaboration with Kirkpatrick et al. (2017), we propose a novel bioinformatics pipeline that can be used to identify chemical compounds found in plants or other organisms that have a deleterious effect on pathogens or cancer cell lines.

Explanatory data is collected through liquid chromatograpy - mass spectrometry (LC-MS). In brief, this involves using liquid chromatograpy to separate out the compounds that exist in a given plant tissue into several dozen solutions known as fractions. A given compound is observed (has a nonzero quantity) over only a few fractions, so that compounds are better differentiated when performing downstream statistical analyses. The mass spectrometry phase is used to actually measure the quantity of the various compounds present in each of the fractions. The outcome variables are obtained by introducing solution from a given fraction into a Petri dish with the pathogen or cancer cell line of interest, and observing the change in the quantity of the pathogen or cell line after a fixed amount of time.

The size of the data is a primary challenge for this setting. The explanatory data consists of over 10,000 readings from the mass spectrometer for each observation. The number of observations is determined by the number of replications for each of the fractions of interest (i.e. the fractions in which a decrease in the pathogen or cancer cell line is observed). Performing these experiments is expensive and time-consuming, and for our real data analyses we had about 15 observations for a given outcome of interest. Given this scarcity of data, effective screening of the data is of critical importance to us. Furthermore, multiple observations read by the mass spectrometer can belong to a single compound, so we developed a procedure to recover the underlying compounds. Finally, after compound recovery and candidate compound screening is performed, we propose a simple method of ranking the candidate compounds that works effectively in this setting of acute data scarcity.

1.3 Fertility models

The study of human fertility is an effort to understand the mechanisms by which conception is achieved, and has applications for couples trying to conceive, users of natural family-planning methods, and clinicians trying to estimate probabilities of pregnancy. The use of probability models in this field has shown to provide valuable insights into the predictors of the probabilities of pregnancy. The study of fertility through the use of probability models can be divided into two paradigms based terms time-granularity: menstrual cycle-level models and day-level models. Which class of models to use can be selected based on the analysis goals and availability of data.

1.3.1 Cycle-level time scale

A common approach to modeling fertility data is to consider each menstrual cycle as the time-scale, and use a time-to-event strategy to model pregnancy status. The most popular model for this setting is the discrete-time proportional hazards model proposed in Cox (1972). The model is described as follows.

1.3.1.1 Discrete-time proportional hazards model

Let T be a discrete random variable with support $t_1 < t_2 < \ldots$, and let us write the pmf of T as $f(t_j) = f_j = \mathbb{P}\left(T = t_j\right)$. Then the survival function denoted as $S(t_j)$ is given by

$$S(t_j) = S_j = \mathbb{P}\left(T \geq t_j\right) = \sum_{k=j}^{\infty} f_j.$$

Next, let us define the hazard at time t_j, denoted as $\lambda(t_j)$, to be the conditional probability of pregnancy given that no pregnancy has occurred up to that point, then it follows that

$$\lambda(t_j|\, \boldsymbol{x}_{ij}) = \lambda_j = \mathbb{P}\left(T = t_j \,|\, T \geq t_j, \boldsymbol{x}_{ij}\right) = \frac{f_j}{S_j}.$$

The discrete-time proportional hazards model is defined as

$$\frac{\lambda(t_j|\boldsymbol{x}_{ij})}{1 - \lambda(t_j|\boldsymbol{x}_{ij})} = \frac{\lambda_0(t_j)}{1 - \lambda_0(t)} \exp\{\boldsymbol{x}_{ij}^T \boldsymbol{\beta}\}.$$

Estimation of the regression coefficients then proceeds through maximizing the partial likelihood.

1.3.2 Day-level time scale

One of the primary methods used to estimate to fit a model based on day-specific data is based on Schwartz et al. (1980). In this model the probability of pregnancy for a cycle, given that a pregnancy has not already occurred in a previous cycle, is of the form

$$1 - \prod_{k=1}^{K} \left(1 - \lambda_k\right)^{X_{ijk}},$$

for subject i, cycle j, and fertile window day k, and where X_{ijk} is an indicator variable taking a value of 1 if sexual intercourse occurred on the corresponding day and 0 otherwise. Under this model, λ_{ijk} has the interpretation of the day-specific probability of conception in cycle j from couple i given that conception has not already occurred, or equivalently, given intercourse only on day k. Estimation of the λ_{ijk}'s proceeds through maximum likelihood estimation.

The Schwartz *et al.* model has the ability to estimate the day-specific probability of conception under the assumption that probabilities are the same across couples and cycles, which seems unlikely in actuality. Furthermore, scientific interest often lies not just in estimating these probabilities, but identifying predictors that may be protective or promotive to achieving pregnancy. To address these issues, Dunson and Stanford (2005) proposed hierarchical Bayesian model with a regression term that allows for the incorporation of predictors into the probability model. Their model is specified as follows. Define

Y_{ij} an indicator of conception for woman i, cycle j,

V_{ijk} an indicator of conception for woman i, cycle j, day k,

X_{ijk} an indicator of intercourse for woman i, cycle j, day k,

and let

$\boldsymbol{u}_{ijk}$ a covariate vector of length q for woman i, cycle j, day k,

$\boldsymbol{\beta}$ a vector of length q of regression coefficients,

ξ_i woman-specific random effect.

Then the model is given by

$$\begin{aligned}
\mathbb{P}\left(Y_{ij} = 1 \,\middle|\, \xi_i, \boldsymbol{X}_{ij}, \boldsymbol{U}_{ij}\right) &= 1 - \prod_{k=1}^{K} (1 - \lambda_{ijk})^{X_{ijk}} \\
\lambda_{ijk} &= 1 - \exp\left\{-\xi_i \exp\left(\boldsymbol{u}'_{ijk}\boldsymbol{\beta}\right)\right\} \\
\xi_i &\sim \text{Gamma}\,(\phi, \phi),
\end{aligned}$$

1.3.2.1 Novel research: extending the day-specific probabilities model

The model of Dunson and Stanford (2005), while arguably the gold standard for this domain, has several limitations that we try to address in this research. The first is that the model only permits categorical predictor variables. While categorization of certain variables such as age is a widely used technique in this field, there are other variables of interest such as for example inhibin levels, a protein complex that can be used as a proxy for ovarian reserve, for which established categorization cutpoints may not be known. When this is the case or when a potential feature may be more naturally represented as a continuous variable then having a means to do so is desirable. We modify the approximation of the posterior density in our research so as to allow for this type of data.

A second potential issue is that missing data is not accounted for in the model, and as a result data with missing values is typically omitted from data analyses. Of course, as with any domain this is an undesirable practice, and in particular, fertility studies are quite expensive to set up and administer so data is at a premium. But what may not be apparently obvious is that due to the structure of the data, missing a single variable value for an observation in some sense requires multiple observations to be thrown out. The reason for this is that observations are collected on a day-level time scale, but the pregnancy status outcome variable occurs on a cycle-level time scale. Thus to include a cycle in the model, you cannot have missingness in any of the days that occur during the cycle. As a result, a small proportion of missingness in the data can cause a large amount of loss of data in an analysis. In order to account prevent this effect, we extend the model to explictely include the missing values which allows for posterior inference to include data with missingness.

A third issue that the known true unimodal shape of the fertile window day probabilities is not explicitly accounted for. What this means, is that based on subject-matter biological knowledge, researchers in the field expect that the probability of conception during a given cycle there is a time of peak fertility, and that when considered as a function of time fertility is monotone decreasing as the amount of time increases either before or after this peak. While sufficiently powered analyses should be able to infer this structure, if the model itself can somehow take into account for this structure, then doing so can potentially improve the informativeness of the posterior density. We propose two ways of incorporating this information into the model, each of which may be preferable in different circumstances.

CHAPTER 2: COMPOSITE QUANTILE-BASED CLASSIFIERS

2.1 Quantile classifiers for univariate data

The quantile classifier and the empirically optimal quantile classifier were introduced in Hennig and Viroli (2016). For completeness, the notation and definitions are reviewed in Section 2.1.1. In Section 2.1.2, we propose an extension to the quantile classifiers that we call *multimodal quantile classifiers* for with settings where there are multiple decision rule boundaries. Then in Section 2.1.3 we present some results regarding the optimality and consistency of the new classifier.Additionally, a new algorithm is proposed in the supplementary materials that obtains the decision rule for the empirically optimal quantile classifier in quadratic time.

2.1.1 The quantile classifier

In order to present the quantile classifier, we begin by introducing some terminology originally defined in Hennig and Viroli (2016). Consider the check loss (also known as the quantile loss) function defined as

$$\rho_\theta(u) = \mathbb{1}\left(u > 0\right)\theta\, u + \mathbb{1}\left(u \leq 0\right)(1-\theta)\,(-u) = u\left[\theta - \mathbb{1}\left(u \leq 0\right)\right],$$

for some choice of quantile level $\theta \in (0,1)$. Next, consider two populations Π_0 and Π_1 with corresponding distribution functions F_0 and F_1 on $\mathbb{R}$. The quantile distance of a point z to the population's θ-th quantile is defined as

$$\Phi_i(z,\theta) = \rho_\theta\Big(z - F_i^{-1}(\theta)\Big), \qquad i = 0, 1.$$

It will also be convenient later to define for fixed θ,

$$F_{(0)}^{-1}(\theta) = \min\Big\{F_0^{-1}(\theta),\ F_1^{-1}(\theta)\Big\} \qquad \text{and} \qquad F_{(1)}^{-1}(\theta) = \max\Big\{F_0^{-1}(\theta),\ F_1^{-1}(\theta)\Big\}.$$

Next, the difference of the quantile distances of a point z to the populations' θ-th quantiles is defined as

$$\Lambda\left(z,\theta\right) = \Phi_1(z,\theta) - \Phi_0(z,\theta).$$

With these definitions in hand, the θ-th quantile classifier is defined as follows.

$$\text{For an observation } z\text{, classify to:} \quad \begin{cases} \Pi_0, & \text{if} \quad \Lambda\left(z,\theta\right) > 0, \\ \Pi_1, & \text{otherwise.} \end{cases} \tag{2.1}$$

It can also be shown that the decision rule boundary resulting from the classifier in (2.1) is given by the value of $\theta F_{(0)}^{-1}(\theta) + (1-\theta)\, F_{(1)}^{-1}(\theta)$.

Finally, let Z be a random variable with a prior probability π_0 of being a member of population Π_0, and $\pi_1 = 1 - \pi_0$ the prior probability of being a member of population Π_1. Then the probability of correctly classifying an observed realization of Z by the θ-th quantile classifier is given by

$$\Psi(\theta) = \pi_0 \int \mathbb{1}\left(\Lambda(z,\theta) > 0\right) dP_0(z) + \pi_1 \int \mathbb{1}\left(\Lambda(z,\theta) \leq 0\right) dP_1(z). \tag{2.2}$$

It was shown in Hennig and Viroli (2016) that the expression given in (2.2) has the appealing property that for the optimal choice of quantile level θ and when the population distribution of one of the classes stochastically dominates the other, the classification rate is equal to that of the Bayes classifier. In that paper they call the quantile classifier with the optimal quantile level θ the optimal quantile classifier. Since in practice the optimal choice of θ is unknown, this necessitates the empirically optimal quantile classifier, which is presented next.

Let $x_1, \ldots, x_n$ be samples from a population with a probability density on $\mathbb{R}$. Then we define the θ-th empirical quantile for the population to be given by any minimizer of the sample version of the expected check loss distance defined by

$$\hat{F}^{-1}(\theta) = \arg\min_q \left\{ \theta \sum_{x_i > q} |x_i - q| \; + \; (1-\theta) \sum_{x_i \leq q} |x_i - q| \right\}. \tag{2.3}$$

In the supplementary materials we show that a solution to equation (2.3) providing the θ-th empirical quantile for $x_1, \ldots, x_n$ is given by $\lceil n\theta \rceil$-th largest value of the x_i.

Next, let n_0 and n_1 be the number of observations from Π_0, and Π_1, respectively, and let $n = n_1 + n_2$. Having defined the θ-th empirical quantile for the population, it follows that the empirical difference of the quantile distances of a point z to the population's θ-th quantiles is defined as

$$\Lambda_n(z, \theta) = \rho_\theta\Big(z - \hat{F}_{1n_1}^{-1}(\theta)\Big) - \rho_\theta\Big(z - \hat{F}_{0n_0}^{-1}(\theta)\Big).$$

Then we further define the observed rate of correct classification for the θ-th quantile as

$$\begin{aligned}\Psi_n(\theta) &= \frac{n_0}{n_0+n_1}\left[\frac{1}{n_0}\sum_{x_i \in \Pi_0} \mathbb{1}\Big(\Lambda_n(x_i,\theta) > 0\Big)\right] + \frac{n_1}{n_0+n_1}\left[\frac{1}{n_1}\sum_{x_i \in \Pi_1} \mathbb{1}\Big(\Lambda_n(x_i,\theta) \le 0\Big)\right] \\ &= \frac{1}{n}\left[\sum_{x_i \in \Pi_0} \mathbb{1}\Big(\Lambda_n(x_i,\theta) > 0\Big) + \sum_{x_i \in \Pi_1} \mathbb{1}\Big(\Lambda_n(x_i,\theta) \le 0\Big)\right].\end{aligned}$$

Finally, define the empirically optimal quantile classifier to be any solution to the equation

$$\hat{\theta}_n = \underset{\theta \in T}{\arg\max}\, \Psi_n(\theta), \tag{2.4}$$

where $T = [\delta, 1-\delta]$ for some small positive constant δ. The restriction of quantile levels to the set T is to guarantee at least one optimal value of θ for the theoretical quantile classifier.

2.1.2 The multimodal quantile classifier

We discussed in Section 2.1.1 that for a fixed choice of θ, the quantile classifier has a single decision rule boundary. As a result the classifier may suffer from loss of efficiency in settings where the Bayes rule consists of multiple boundaries. To mitigate this drawback, we propose an extension to the quantile classifier that we call the *multimodal quantile classifier*.

The way that the multimodal quantile classifier works is the following. The populations' domains are partitioned by splitting the domain at any point x where $F_0(x) = F_1(x)$. Then the usual quantile classifier is used to classify the points based only on the subset of the distributions' support that belong to the interval created by the partitioning. See Figure 2.1 for the partitioning by the multimodal quantile classifier for an example setting.

In more detail, suppose that there are a finite number of points $x_1 < \cdots < x_t$ such that $F_0(x) = F_1(x)$ for $x \in \{x_1, \ldots, x_t\}$ and that $F_0(x) \neq F_1(x)$ otherwise, and let us denote $x_0 = \inf\{x : f_0(x) > 0\}$ and $x_{t+1} = \sup\{x : f_0(x) > 0\}$. Then we create intervals $(x_{k-1}, x_k], k = 1, \ldots t+1$ and classify any new point by comparing the quantile distance for the point using an interval-specific quantile level. Formally, the $(t+1)$-interval multimodal quantile classifier is defined as follows.

$$\text{For an observation } z\text{, classify to:} \quad \begin{cases} \Pi_0, & \text{if } \sum_{k=1}^{t+1} \mathbb{1}\big(z \in (x_{k-1}, x_k]\big)\, \Lambda(z, \theta_k) > 0, \\ \Pi_1, & \text{otherwise.} \end{cases} \tag{2.5}$$

Next, the empirically optimal multimodal quantile classifier is constructed as follows, using a two-step process. For the first step in the process, we start by letting $\hat{x}_1, \ldots, \hat{x}_t$ be the values such that $\hat{F}_0(\hat{x}_k) = \hat{F}_1(\hat{x}_k),\ k = 1, \ldots, t$. In practice, it is possible that there will be intervals of points that will satisfy this equation, and in such cases we pick just a single point within each interval such as, for example the midpoint. We further let $x_0 = -\infty$ and $x_{t+1} = \infty$.

Once we have these estimated cutpoints in hand, then the second step is to divide the domain into $t+1$ subproblems, and to find the empirically optimal quantile classifier for each subproblem. In more detail, let us denote the empirical classification rate for a subproblem on the support $(a, b]$ as

$$\Psi_{n,(a,b]}(\theta) = \frac{1}{\sum_{i=1}^{n} \mathbb{1}\big(x_i \in (a, b]\big)} \left[\sum_{\substack{x_i \in \Pi_0 \\ x_i \in (a,b]}} \mathbb{1}\Big(\Lambda_n(x_i, \theta) > 0\Big) + \sum_{\substack{x_i \in \Pi_1 \\ x_i \in (a,b]}} \mathbb{1}\Big(\Lambda_n(x_i, \theta) \leq 0\Big) \right].$$

Then we define the empirically optimal quantile classifier for the subproblem restricted to the support $(a, b]$ to be any solution to the equation

$$\hat{\theta}_{n,(a,b]} = \arg\max_{\theta \in T} \Psi_{n,(a,b]}(\theta),$$

where $T = [\delta, 1-\delta]$ for some small positive constant δ. Then the classification rule is defined as follows.

$$\text{For an observation } z\text{, classify to:} \quad \begin{cases} \Pi_0, & \text{if } \sum_{k=1}^{t+1} \mathbb{1}\big(z \in (\hat{x}_{k-1}, \hat{x}_k]\big)\, \Lambda_n(z, \hat{\theta}_{n,\,(\hat{x}_{k-1}, \hat{x}_k]}) > 0, \\ \Pi_1, & \text{otherwise.} \end{cases}$$

We note that this approach is a more general version of regular quantile classifiers, since it reduces to the original form when there are no points x at which $F_0(x) = F_1(x)$. It also has the advantage of keeping estimation simple, since the only new requirement is estimation of F_0 and F_1, which is something that is essentially already needed since we are previously estimating F_0^{-1} and F_1^{-1}. This leads to an efficient classifier for many of the scenarios that we would expect to encounter in practice.

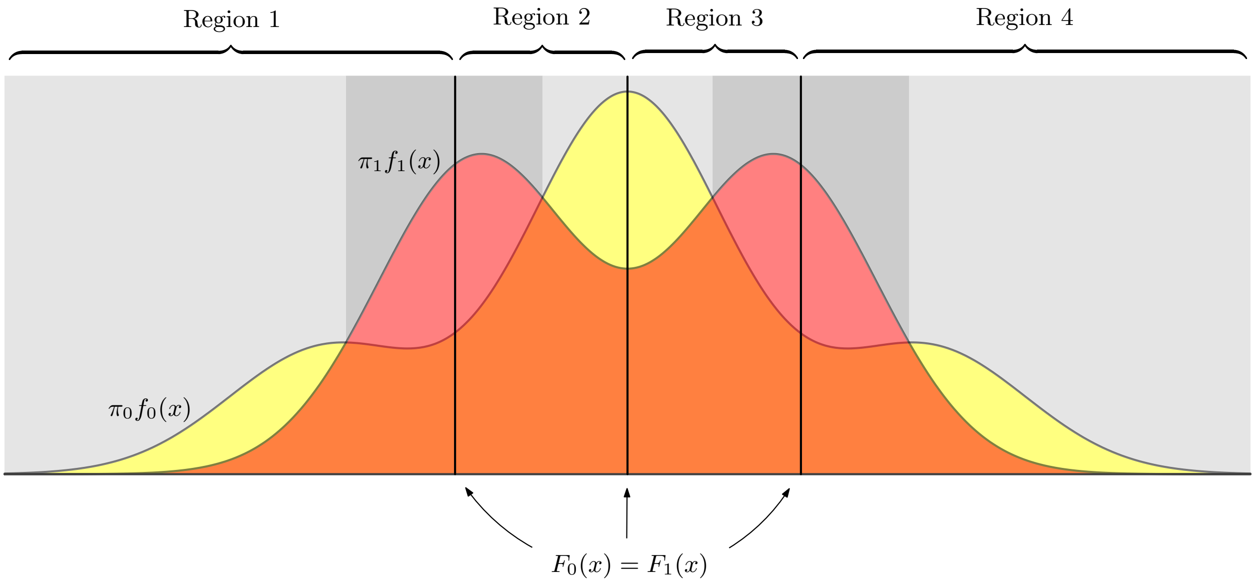


Figure 2.1: The multimodal quantile classifier partitioning for a pair of example populations. In this example there are three values of x for which the distributions functions $F_0(x) = F_1(x)$ are equal, resulting in a partitioning into four regions. The alternating gray regions show the classification regions for the optimal multimodal quantile classifier.

2.1.3 Optimality and consistency of the multimodal quantile classifier

In this section, we start with an alternative convergence result for the empirically optimal quantile classifier. Let us define a variant of Ψ_n which we call $\Psi_n^\dagger$, by

$$\Psi_n^\dagger(\theta) = \pi_0 \int \mathbb{1}\left(\Lambda_n(z,\theta) > 0\right) dP_0(z) + \pi_1 \int \mathbb{1}\left(\Lambda_n(z,\theta) \leq 0\right) dP_1(z).$$

We note that $\Psi_n^\dagger(\hat{\theta}_n)$ is the classification rate for the empirically optimal quantile classifier. Let δ be a small positive constant, and consider the following assumptions.

Assumption 1. F_i^{-1} is a continuous function of θ, $i = 0, 1$.

Assumption 2. $\mathbb{P}\left(\Lambda(Z,\theta) = 0\right) = 0$ for all $\theta \in [\delta, 1-\delta]$.

Theorem 1. *Let $\tilde{\theta}$ be a solution to $\tilde{\theta} = \arg\max_{\theta \in T} \Psi(\theta)$, and let $\mathcal{R}$ be the Bayes rule classification rate for populations Π_0 and Π_1. Then under Assumptions 1 and 2, it follows that $\Psi_n^{\dagger}(\hat{\theta}_n) \xrightarrow{p} \mathcal{R}$.*

Having established this convergence result for the usual quantile classifier, we now we turn our attention to the multimodal quantile classifier.

Theorem 2. *Consider two populations Π_0 and Π_1 with corresponding distribution functions F_0 and F_1 and density functions f_0 and f_1 such that f_0 and f_1 are nonzero on the same domain, and further suppose that Assumptions 1 and 2 hold. Let Z be a random variable with a prior probability π_0 of being a member of population Π_0, and $\pi_1 = 1 - \pi_0$ the prior probability of being a member of population Π_1.*

Next, suppose that there are a finite number of points $x_1 < \cdots < x_t$ such that $F_0(x) = F_1(x)$ for $x \in \{x_1, \ldots, x_t\}$ and that $F_0(x) \neq F_1(x)$ otherwise, and let us denote $x_0 = \inf\{x : f_0(x) > 0\}$ and $x_{t+1} = \sup\{x : f_0(x) > 0\}$. Further assume that for every interval $(x_k,\, x_{k+1})$, that there is exactly one point $z_k^* \in (x_k,\, x_{k+1})$ such that $\pi_0 f_0(z_k^*) = \pi_1 f_1(z_k^*)$, and that $\pi_0 f_0(z) < \pi_1 f_1(z)$ for z on one side of z_k^* and $\pi_0 f_0(z) > \pi_1 f_1(z)$ for z on the other side of z_k^* in the interval. Then the multimodal quantile classifier for an observed realization of Z using the optimal choice of quantile levels achieves the Bayes error rate.

This theorem describes the settings in which the multimodal quantile classifier can be fully efficient. Since, the classifier is limited to a single decision rule boundary for every interval (x_k, x_{k+1}), this will not be sufficient for settings where multiple boundaries may be needed within the interval. On the other hand, this kind of situation will tend to occur for small regions of the data, since larger regions would cause additional points where $F_0(x) = F_1(x)$. Next, we consider the consistency of the empirical version of the classifier.

Theorem 3. *Consider the setup of Theorem 2. Let $\Psi_n^{\dagger}$ denote the classification rate of the empirically optimal multimodal quantile classifier, and let $\mathcal{R}$ denote the Bayes rule classification rate. Then under Assumptions 1 and 2 from Theorem 1, it follows that $\Psi_n^{\dagger} \xrightarrow{p} \mathcal{R}$.*

Thus we see that in settings where the multimodal classifier can be fully efficient, the classification rate of the empirical multimodal classifier converges to the Bayes rule classification rate.

2.2 Quantile-based classifiers for multivariate data

Having characterized the quantile classifier, which is by nature inherently univariate, our goal is to construct a classification method for multivariate settings. A classifier for multivariate data was proposed in

Hennig and Viroli (2016) that bases the classification rule on the sum of the component-wise differences of the quantile distances. In this section we propose a classifier that generalizes this approach in several directions by allowing for more flexible choices of quantile levels, and basing the classifier on a linear combination of the differences of the quantile distances.

We begin this section by extending some of the notation from the previous section to the multivariate setting. First, consider two populations Π_0 and Π_1 with corresponding distribution functions F_0 and F_1 each on $\mathbb{R}^p$, and further let F_{0j} and F_{1j} denote the marginal distribution functions for the j-th variable with respect to F_0 and F_1, $j = 1, \ldots, p$. Then we denote the quantile distance of a point $z \in \mathbb{R}$ to the θ-th quantile of a population's j-th component as

$$\Phi_{ij}(z, \theta) = \rho_\theta\Big(z - F_{ij}^{-1}(\theta)\Big), \qquad i = 0, 1, \quad j = 1, \ldots, p.$$

Using these definitions, the difference of the quantile distances of a point z to the populations' θ-th quantile of the j-th component is defined as

$$\Lambda_j(z, \theta) = \Phi_{1j}(z, \theta) - \Phi_{0j}(z, \theta), \qquad j = 1, \ldots, p.$$

The main idea of the composite quantile-based classifiers proposed in the next section is to aggregate the discriminatory information contained in each of the Λ_j's to construct a classification method for data with multivariate features.

2.2.1 Composite quantile-based classifiers

In this section we propose the family of composite quantile-based classifiers. We begin by motivating the the form of the classifiers. Consider a point $\boldsymbol{z} = (z_1, \ldots, z_p) \in \mathbb{R}^p$ and quantile levels $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p) \in (0, 1)^p$. If the difference of the quantile distances of the j-th component $\Lambda_j(z_j, \theta_j)$ is positive, this provides some descriminatory information indicating that z_j is closer to the θ_j-th quantile of the j-th component of population Π_0 than it is to the θ_j-th quantile of the j-th component of population Π_1, and conversely if $\Lambda_j(z_j, \theta_j)$ is negative. Furthermore, differences of larger magnitudes provide stronger information than those with smaller magnitudes. With these properties in mind we propose using a linear combination of the

$\Lambda_j(z_j, \theta_j)$'s as follows. We call this family of classifiers *composite quantile-based classifiers*.

$$\text{For an observation } \boldsymbol{z}\text{, classify to: } \begin{cases} \Pi_0, & \alpha_0 + \sum_{j=1}^{p} \alpha_j \, \Lambda_j(z_j, \theta_j) > 0, \\ \Pi_1, & \text{otherwise,} \end{cases} \tag{2.6}$$

for $\boldsymbol{\alpha} = (\alpha_0, \ldots, \alpha_p) \in \mathbb{R}^{p+1}$. By introducing the coefficients to the classification model, this in principle provides the ability to both handle the differences in scale of the features and to weight the variables as a function of their discriminatory power.

Since the values of the Λ_j's are a measure of closeness to the θ-th quantile of the parent populations, it may seem strange that the coefficients are allowed to be negative since a negative value makes the classifier less likely to select the population which it is closer to. However, this is an important characteristic of the classifier since it allows the aggregation step to most flexibly model dependencies between the features in the data. For example, when classifying between two Gaussian distributions with a common covariance, the decision rule boundary is linear with the slope determined by the means and covariance of the distributions. Thus, by allowing for negative linear combination coefficients, the classifier can well approximate such a decision rule regardless of the correlation structure in the data.

Finally, the coefficient α_0 can be interpreted as an offset to classify an observation when the linear combination of the differences of the quantile distances is small. This is useful when the prior probability of an observation being drawn from one class is higher than being drawn for the other. When this is the case, we can require a higher threshold of closeness for an observation to a population that has a low prior probability before we decide to classify the observation to that population.

2.2.1.1 Properties of composite quantile-based classifiers

Having presented the theoretical form of the composite quantile-based classifiers, we now consider the consistency of the empirical version as well as the form of the classifier decision rule boundary. Firstly, let us denote $\Psi(\boldsymbol{\theta})$ as the classification rate for composite quantile-based classifiers given by

$$\begin{aligned} \Psi(\boldsymbol{\theta}) = \pi_0 \int \mathbb{1} \left(\alpha_0 + \sum_{j=1}^{p} \alpha_j \, \Lambda_j(z_j, \theta_j) > 0 \right) dP_0(\boldsymbol{z}) \\ + \pi_1 \int \mathbb{1} \left(\alpha_0 + \sum_{j=1}^{p} \alpha_j \, \Lambda_j(z_j, \theta_j) \leq 0 \right) dP_1(\boldsymbol{z}). \end{aligned}$$

For the theorem below we will need the following assumptions. Let δ be a small positive constant.

Assumption 1. F_{ij}^{-1} is a continuous function of $\theta,\ i = 0, 1,\ j = 1, \dots, p$.

Assumption 2. $\mathbb{P}\left(\alpha_0 + \sum_{j=1}^{p} \alpha_j\, \Lambda_j(Z_j, \theta_j) = 0\right) = 0$ for all $\theta \in [\delta, 1 - \delta]$.

Assumption 3. There is a unique $\tilde{\theta}_j$ that satisfies $\tilde{\theta}_j = \arg\max_{\theta \in T} \Psi_j(\theta), j = 1, \dots, p$.

Theorem 4. *Denote $\tilde{\boldsymbol{\theta}} = \left(\tilde{\theta}_1, \dots, \tilde{\theta}_P\right)$. Then under assumptions 1-3 it follows that $\Psi(\hat{\boldsymbol{\theta}}_n) \xrightarrow{p} \Psi(\tilde{\boldsymbol{\theta}})$.*

This result shows that the classification rate of the empirical version of composite quantile-based classifiers converges to that of the composite quantile-based classifier using the component-wise most discriminatory choices of quantile levels.

Next, we consider the decision rule form of composite quantile-based classifiers. Recall that the composite quantile-based family of classifiers is defined as expression (2.6). From this expression we see that the decision rule boundary is given by

$$\alpha_0 + \sum_{j=1}^{p} \alpha_j\, \Lambda_j(z_j, \theta_j) = 0. \tag{2.7}$$

As an aside, we note that the set of $\boldsymbol{z}$ that satisfy (2.7) may be empty for certain choices of $(\alpha_0, \dots, \alpha_p, \theta_1, \dots, \theta_p)$, which corresponds to a decision rule that exclusively classifies observations to one of the two classes. If the solution set is nonempty, then we can show that the decision rule boundary is a piecewise linear function with a particularly simple form, as is discussed in what follows. Let us define the component-wise quantile classifier decision boundary τ_j to be the unique point satisfying $\Lambda_j(\tau_j, \theta_j) = 0,\ j = 1, \dots, p$. We can assume without loss of generality that τ_j is unique because it is only non-unique when $\Lambda(\,\cdot\,, \theta_j)$ is identically zero, in which case the problem is effectively reduced to $p - 1$ dimensions. It follows then that

$$\operatorname{sign}\left(F_{1j}^{-1}(\theta_j) - F_{0j}^{-1}(\theta_j)\right) \Lambda_j(z_j, \theta_j) = \begin{cases} \tau_j - F_{(0)j}^{-1}(\theta_j), & z_j \le F_{(0)j}^{-1}(\theta_j), \\ \tau_j - z_j, & F_{(0)j}^{-1}(\theta_j) < z_j < F_{(1)j}^{-1}(\theta_j), \\ \tau_j - F_{(1)j}^{-1}(\theta_j), & \text{otherwise}. \end{cases} \tag{2.8}$$

The expression in (2.8) leads to the decision rule form of composite quantile-based classifiers. Conceptually, we can divide $\mathbb{R}^p$ into 3^p hypercubes so that the decision rule boundary set is either empty or is an affine set within each hypercube.

Proposition 5. *Define* $\alpha_j^* = \alpha_j \operatorname{sign}\Big(F_1^{-1}(\theta_j) - F_0^{-1}(\theta_j)\Big)$*, and let* L*,* M*, and* U *(for lower, middle and upper) be sets that partition the set of indices* $\{1, \ldots, p\}$*. Define the hypercube* $\mathcal{H}_{L,M,U}$ *to be the set given by*

$$\mathcal{H}_{L,M,U} = \prod_{\ell \in L} \big(-\infty,\ F_{(0),\ell}^{-1}(\theta_\ell)\big) \prod_{m \in M} \big[F_{(0),m}^{-1}(\theta_m),\ F_{(1),m}^{-1}(\theta_m)\big] \prod_{u \in U} \big(F_{(1),u}^{-1}(\theta_u),\ \infty\big), \tag{2.9}$$

where in this context the products denote the usual Cartesian product. Then the decision rule boundary for composite quantile-based classifiers on the domain $\mathcal{H}_{L,M,U}$ *is the possibly empty set given by*

$$\mathcal{D}_{L,M,U} = \Bigg\{ \boldsymbol{z} \in \mathcal{H}_{L,M,U} : \Bigg[\alpha_0 + \sum_{\ell \in L} \alpha_\ell^* \Big(\tau_\ell - F_{(0),\ell}^{-1}(\theta_\ell)\Big) + \sum_{u \in U} \alpha_u^* \Big(\tau_u - F_{(1),u}^{-1}(\theta_u)\Big) \Bigg] + \sum_{m \in M} \alpha_m^* (z_m - \tau_m) \ = \ 0 \Bigg\}.$$

Furthermore, the decision rule boundary is a continuous function of $\boldsymbol{z}$*.*

2.2.1.2 Augmented composite quantile-based classifiers

The general approach of building a classifier based on the one-at-a-time comparisons of the marginal densities can be a useful strategy for reducing the complexity of a problem in a high-dimensional setting. However, considering only the marginal distributions can result in deletion of features that could otherwise provide discriminative information when considered jointly. In the case of composite quantile-based classifiers, deletion will occur whenever the marginal distribution functions are equal for the specified quantile level.

In order to prevent such a loss of information, we propose a variant of the classifiers called *augmented composite quantile-based classifiers*. In this variant, the transformed features upon which the decision rule boundary are constructed are augmented by the original features. In other words, the decision rule for an observation $\boldsymbol{z}$ is based upon

$$\alpha_0 + \alpha_1 \Lambda_1(z_1, \theta_1) + \cdots + \alpha_p \Lambda_p(z_p, \theta_p) + \alpha_{p+1} z_1 + \cdots + \alpha_{2p} z_p.$$

This allows the classification rule to be constructed using the full multivariate information. Additionally, other forms of correlation can be encoded into the data by including interaction terms or other forms of feature expansion.

Including the original features can be helpful whenever deletion may occur or when the decision rule boundary is approximately linear in the original features. Numerical results for the two variants are compared in Section 2.3, and in the supplementary materials this phenomenon is investigated further by comparing numerical results for a setting that is subject to feature deletion.

2.2.1.3 Multimodal composite quantile-based classifiers

Extending the composite quantile-based classifiers to use the multimodal quantile classifier is straightforward. The main issue that needs to be resolved is how to define the Λ_j's, since there are multiple quantile levels to consider within a feature. Suppose that features' marginal support is divided into $t+1$ regions, where t is the number of values of x where $F_0(x) = F_1(x)$. Let us denote these regions as $\mathcal{A}_{j1}, \ldots, \mathcal{A}_{j,t+1}$ ordered so that $x \in \mathcal{A}_{jk}$ and $x' \in \mathcal{A}_{jk'}$ for $k < k'$ implies that $x < x'$. Then for a given region $\mathcal{A}_{jk}$, a quantile level θ_{jk} is chosen with a resulting decision rule boundary that we call τ_{jk} such that $\tau_{jk} \in \mathcal{A}_{jk}$.

For values of z with $z < \tau_{j1}$ or $\tau_{j,t+1} < z$, it is natural to define the difference of the quantile distances with respect to θ_{j1} or $\theta_{j,t+1}$ (c.f. Figure 2.1). For z with $\tau_k < z < \tau_{k+1}$, then there are two potential quantiles to base a distance upon. We propose using a combination of both quantile levels based on the point's relative distances to the regions' decision rule boundary. Formally, we define the transformation by

$$z_j^* = \begin{cases} \Lambda_j(z_j, \theta_{j1}), & z_j < \tau_{j1}, \\ \frac{\tau_{k+1} - z_j}{\tau_{k+1} - \tau_k} \Lambda_j(z_j, \theta_{jk}) + \frac{z_j - \tau_k}{\tau_{k+1} - \tau_k} \Lambda_j(z_j, \theta_{j,k+1}), & \tau_k \leq z_j < \tau_{k+1}, k = 1, \ \ldots, t, \\ \Lambda_j(z_j, \theta_{j,t+1}), & \tau_{j,t+1} \leq z_j. \end{cases}$$

We note that this form of z_j^* results in a function that is continuous and piecewise linear in z_j. When the assumptions of Theorem 2 are met and for optimal choices of quantile levels, then $z_j^* < 0$ when $\pi_0 f_0(x) > \pi_1 f_1(x)$, $z_j^* = 0$ when $\pi_0 f_0(x) = \pi_1 f_1(x)$, and $z_j^* > 0$ when $\pi_0 f_0(x) < \pi_1 f_1(x)$. Thus this data transformation maintains many of the characteristics that are seen in the usual composite quantile-based classifiers.

2.2.2 Classifier model selection

Having introduced a family of classifiers in Section 2.2.1, we are in need of a method by which to select a particular model based upon the data at hand. This requires selection of both a vector of component-wise quantile levels, and a vector of component coefficients. To perform model selection, we propose a two-step process as described in Sections 2.2.2.1 and 2.2.2.2. Additionally, in Section 2.2.4 we present an alternative algorithm that avoids splitting data at the cost of expanding the feature space.

2.2.2.1 Choice of quantile levels

As stated before, the fundamental motivation for the family of classifiers proposed in this paper is the result that for univariate data and under some assumptions, the decision rule based upon the distances of an observation to the corresponding within-class quantiles for the optimal choice of quantile levels is the Bayes rule. Thus, we propose using these most powerful choices of quantile level with respect to each feature's discriminatory ability, taken one-at-a-time. This statement is developed more precisely below.

To begin, let us suppose that we have p-dimensional observations $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, and we let x_{ij} denote the j-th component of $\boldsymbol{x}_i$; $i = 1, \ldots, n,\ j = 1, \ldots, p$. Then the empirical quantile for the j-th feature of the i-th population is defined to be any solution satisfying

$$\hat{F}_{ijn}^{-1}(\theta) = \arg\min_q \left\{ \theta \sum_{\substack{\{\ell \colon \boldsymbol{x}_\ell \in \Pi_i \\ x_{\ell j} > q\}}} |x_{\ell j} - q| \ + \ (1-\theta) \sum_{\substack{\{\ell \colon \boldsymbol{x}_\ell \in \Pi_i \\ x_{\ell j} \leq q\}}} |x_{\ell j} - q| \right\}.$$

Having defined the θ-th empirical quantile for the j-th feature of the i-th population, it follows that the empirical difference of the quantile distances of a point $z \in \mathbb{R}$ to the population's θ-th quantiles of the j-th feature is defined as

$$\Lambda_{jn}(z, \theta) = \rho_\theta\Big(z - \hat{F}_{1jn}^{-1}(\theta)\Big) - \rho_\theta\Big(z - \hat{F}_{0jn}^{-1}(\theta)\Big).$$

Further define the observed rate of correct classification for the θ-th quantile of the j-th feature as

$$\Psi_{jn}(\theta) = \frac{1}{n}\left[\sum_{x_i \in \Pi_0} \mathbb{1}\Big(\Lambda_{jn}(x_{ij}, \theta) > 0\Big) + \sum_{x_i \in \Pi_1} \mathbb{1}\Big(\Lambda_{jn}(x_{ij}, \theta) \leq 0\Big)\right].$$

Finally, define the empirically optimal quantile classifier for the j-th feature to be any solution to the equation

$$\hat{\theta}_{jn} = \arg\max_{\theta \in T} \Psi_{jn}(\theta),$$

for $T = [\delta, 1 - \delta]$ where δ is an arbitrarily small positive constant. Then we propose selecting $\hat{\theta}_{jn}$ as the choice of quantile level, $j = 1, \ldots, p$.

In the case of the multimodal version, we choose a quantile level for each of the regions based on the observed data within the region.

2.2.2.2 Choice of linear combination coefficients

To motivate the choice of linear combination coefficients, we note the following observation. If we write $z_j^* = \Lambda_j(z_j, \theta_j);\ j = 1, \ldots, p$, then we can express the classification rule in equation (2.6) as follows:

$$\text{For an observation } \boldsymbol{z}\text{, classify to: } \begin{cases} \Pi_0, & \alpha_0 + \sum_{j=1}^{p} \alpha_j z_j^* \ > \ 0, \\ \Pi_1, & \text{otherwise.} \end{cases} \tag{2.10}$$

From this form it is readily apparent that the the classification rule decision boundary for the classifier in (2.10) is the same as that of the decision rule boundary obtained from the logistic regression model with $\boldsymbol{z}^* = (z_1^*, \ldots, z_p^*)$ as the predictor variables. From this perspective, it is natural to select the classifier component coefficients by using the logistic regression coefficient estimates obtained from the transformed training data where the transformation is taken to be $x_{ij}^* = \Lambda(x_{ij}, \theta_j)$ for all i, j. Indeed, this is exactly the approach that we propose, using a two step process as follows: first choose quantile levels based on each feature's empirically optimal quantile level in terms of class prediction, and then for these fixed choices of quantile levels use penalized logistic regression on the transformed training data to select a choice of linear combination coefficients. A full presentation of the two-step parameter selection process is provided in Section 2.2.3.

2.2.2.3 Connections to existing methods

Composite quantile-based classifiers share a relationship with a number of existing methods such as quantile-based classifiers, and more generally the class of distance-based classifiers to which quantile-based classifiers belongs. Distance-based classifiers are a type of classifier that assign an observation to a population which it is deemed closer to by some measure of distance. More formally, let $\boldsymbol{\xi}_0$ and $\boldsymbol{\xi}_1$ be p-variate statistics representing populations Π_0 and Π_1, respectively, and further let $d(\cdot,\cdot)$ denote a chosen distance measure. Then, following the notation used in Hennig and Viroli (2016), a distance-based classifier has the following form:

$$\text{For an observation } \boldsymbol{z}\text{, classify to:} \quad \begin{cases} \Pi_0, & \sum_{j=1}^{p} \Big\{ d(z_j, \xi_{1j}) - d(z_j, \xi_{0j}) \Big\} > 0, \\ \Pi_1, & \text{otherwise.} \end{cases}$$

Some examples of distance-based classifiers include nearest centroid classification (e.g. Hastie et al. (2009)), shrunken centroids (Tibshirani et al. (2002), Wang and Zhu (2007)), median-based classifiers (Jörnsten (2004), Ghosh and Chaudhuri (2005), Hall et al. (2012)), and quantile-based classifiers (Hennig and Viroli (2016)). When we restrict the linear combination coefficients so that $\alpha_0 = 0$ and $\alpha_1 = \cdots = \alpha_p > 0$, and also require that $\theta_1 = \cdots = \theta_p = \theta$, then the decision rule is based upon

$$\alpha_0 + \sum_{j=1}^{p} \alpha_j \, \Lambda_j(z_j, \theta_j) \; > \; 0 \qquad \Longleftrightarrow \qquad \sum_{j=1}^{p} \Lambda_j(z_j, \theta) > 0,$$

which is the form of quantile-based classifiers. Quantile-based classifiers belong to the family of distance based classifiers, which can be seen by letting $\boldsymbol{\xi}_i = \Big(F_{i1}^{-1}(\theta), \ldots, F_{ip}^{-1}(\theta)\Big),\ i = 0, 1$, and further letting $d_\theta(z_j, \xi_{ij}) = \rho_\theta(z_j - \xi_{ij})$, in which case

$$\sum_{j=1}^{p} \Lambda_j(z_j, \theta) = \sum_{j=1}^{p} \Big\{ d_\theta(z_j, \xi_{1j}) - d_\theta(z_j, \xi_{0j}) \Big\}.$$

So from this we see that composite quantile-based classifiers cover quantile-based classifiers. Furthermore, although composite quantile-based classifiers are not in general distance based classifiers, we see that a subset of the family of classifiers belongs to the set of distance based classifiers.

Another interesting connection to note is that the proposed choice of quantile levels and variable coefficients from Sections 2.2.2.1 and 2.2.2.2 selects a model from the composite quantile-based classifiers that bears a number of similarities to the FANS classifier proposed in Fan et al. (2016). The FANS classifier starts from the idea that for data with an equal prior probability of being a member from one of two classes with corresponding class conditional densities f_0 and f_1, the Bayes rule decision boundary is given by $\Big\{ \boldsymbol{z} \colon \log \frac{f_0(\boldsymbol{z})}{f_1(\boldsymbol{z})} = 0 \Big\}$. Then, if it is the case that conditional on class membership the features are mutually independent, the Bayes rule decision boundary becomes

$$\mathcal{D} = \left\{ \boldsymbol{z} : \; \log \frac{f_{01}(z_1)}{f_{11}(z_1)} + \cdots + \log \frac{f_{0p}(z_p)}{f_{1p}(z_p)} \; = \; 0 \right\},$$

where f_{ij} corresponds to the class conditional marginal density for the j-th feature of the i-th class. Note that this is the Naive Bayes classifier when $\pi_0 = \pi_1$. The FANS classifier then proposes the following generalization of the decision boundary:

$$\mathcal{D}_{\text{FANS}} = \left\{ \boldsymbol{z} : \; \alpha_0 + \alpha_1 \log \frac{f_{01}(z_1)}{f_{11}(z_1)} + \cdots + \alpha_p \log \frac{f_{0p}(z_p)}{f_{1p}(z_p)} \; = \; 0 \right\}.$$

The classifier then estimates class conditional marginal densities f_{ij} using nonparametric kernel density estimation, and selects the choice of coefficients $\alpha_0, \ldots, \alpha_p$ by using the penalized logistic regression coefficient estimates obtained from the transformed training data where the transformation is taken to be $x^*_{ij} = \log \frac{\hat{f}_{0j}(x_{ij})}{\hat{f}_{1j}(x_{ij})}$ for all i, j.

When compared to distance-based classifiers such as the nearest centroid and shrunken centroid classifiers, composite quantile-based classifiers are similar in that the transformation is based on some notion of distance to a sentinel point of the within-class marginal distributions. However, as discussed in this paper and in Hennig and Viroli (2016), the center of the data is not always the most informative location for such a comparison, and furthermore as discussed in Hall et al. (2009), the use of the Euclidean distance can cause the classifier to suffer from poor performance in certain cases such as in the presence of heavy-tailed data.

When compared to Naive Bayes and FANS, the transformation for composite quantile-based classifiers is similar in that Naive Bayes and FANS map values of x with $f_{0k}(x) > f_{1k}(x)$ to positive values and values of x with $f_{0k}(x) < f_{1k}(x)$ to negative values, while composite quantile-based classifiers similarly map values of x with $\pi_0 f_{0k}(x) > \pi_1 f_{1k}(x)$ to positive values and values of x with $\pi_0 f_{0k}(x) < \pi_1 f_{1k}(x)$ to negative values, under some assumptions. One important difference between the approaches occurs between the estimation required for the methods. The estimation for composite quantile-based classifiers requires estimating $F^{-1}_{0k}(x)$ and $F^{-1}_{1k}(x)$, and selecting an optimal quantile level, all fairly easy tasks. On the other hand, Naive Bayes traditionally requires an a priori assumption on the density of the data, while FANS requires kernel density estimation which is far more difficult to estimate. We will revisit this topic in Section 2.3, where the numerical results indicate that composite quantile-based classifiers perform comparatively better when the quantity of the training data is limited, presumably due in part to these differences in estimation.

One final point that was noted in Fan et al. (2016) and bears repeating here, is that instead of using penalized logistic regression, composite quantile-based classifiers may well use the SVM (linear kernel) or any other linear classifier as a means through which to obtain the transformed feature coefficient values. In this paper, as in Fan et al. (2016), we focus on using penalized logistic regression with the ℓ_1 penalty.

2.2.3 Algorithm for composite quantile-based classifiers

In this section, we present an algorithm used to select a composite quantile-based classifier model. This algorithm is similar to the algorithm proposed in Fan et al. (2016).

We begin with some observations that motivate the form of the algorithm. One issue that needs some consideration is the fact that the proposed choices of quantile levels and linear combination coefficients are based upon a two-step process. Recall from Sections 2.2.2.1 and 2.2.2.2 that the quantile levels are chosen first, with each quantile level based on the optimal level with respect to the feature's ability to predict class membership. Then for the second step, the linear combination coefficients are chosen based upon the transformed data where the transformation is performed with respect to the choices of quantile levels from the first step. Because of this two-step process, the second step is particularly prone to overfitting the model to the training data. To mitigate this undesirable behavior, the data set is split into two parts: one part to train for the choice of quantile levels and corresponding within-class quantiles, and the other part to train for the choice of penalty parameter used to select the linear combination coefficients. The question then is how to split the data. Based on empirical evidence, we suggest randomly splitting the data into evenly-sized partitions. In our experience, this equitable partitioning tends to work well across different sizes and types of data.

While alleviating one problem, splitting the data leads to another concern: namely that different splits of the data can lead to different model choices. To protect against model instability, the data are split not just once but multiple times, and a classification model is procured for each split. The final model is then obtained by averaging the individual models in the following sense. Let f_i be one of the individual models among a total of L data splits. Then a new observation $\boldsymbol{z}$ is classified according to $\frac{1}{L}\sum_{i=1}^{L} f_i(\boldsymbol{z})$ where the decision rule boundary is 0 (furthermore, since the boundary rule is 0 we need not divide by L). We note that the process described here is similar in concept to the ideas presented in bagging Breiman (1996) and random forests Breiman (2001). Another related idea is random-projection ensemble classification Cannings and Samworth (2017), where high-dimensional data are randomly projected into a lower dimension before application by an arbitrary classifier, and then the results are aggregated to construct a decision rule. Random-projection ensemble classification employs data splitting to choose the empirically best projection among a group of projections in an a manner that is similar in spirit to the data splitting procedure proposed here.

The composite quantile-based classifiers algorithm is presented in its entirety in the supplementary materials.

2.2.4 Alternative model selection algorithm: simultaneous selection

One potential downside to the data splitting approach is that we only use one portion of the data at a time to estimate quantiles and select a quantile level, and the other portion to select the linear combination coefficients. To prevent the need for splitting the data, one reviewer of this paper suggested that rather than using a portion of the data to select the quantile levels for each feature, that we instead consider a grid of candidate quantile levels, and include all of these levels in the transformed data. So for example, the decision rule could be based on

$$\alpha_0 + \sum_{k=1}^{q} \sum_{j=1}^{p} \alpha_{jk} \, \Lambda_j\big(z_j, \, k/(q+1)\big),$$

for some positive integer q. Under this scheme, the number of features in the transformed data is q times greater than the number of features in the original data, and we rely on a linear combination coefficients selection algorithm such as penalized logistic regression to find the most important feature and quantile level combinations to use for predictions. In high-dimensional settings, this multiplicative increase in the number of features quickly becomes untenable, and in this case we recommend an additional step of feature screening as is done in Fan and Lv (2008).

We performed a series of simulation studies using this approach and found that it offered competitive performance to the data splitting approach, but were unable to find a clear reason to prefer one approach to the other from a performance standpoint for most settings. One exception occurs for settings where the decision rule boundary is approximately linear, which tend to perform better for the data splitting approach since the original data for the augmented models doesn't get overwhelmed by the multiplicative expansion of the number of features in the transformed data. We show an example of this in Table 9 in the supplementary materials, while in Table 10 we show a more typical example where the different algorithms show very similar levels of prediction accuracy.

2.2.5 Classifier examples

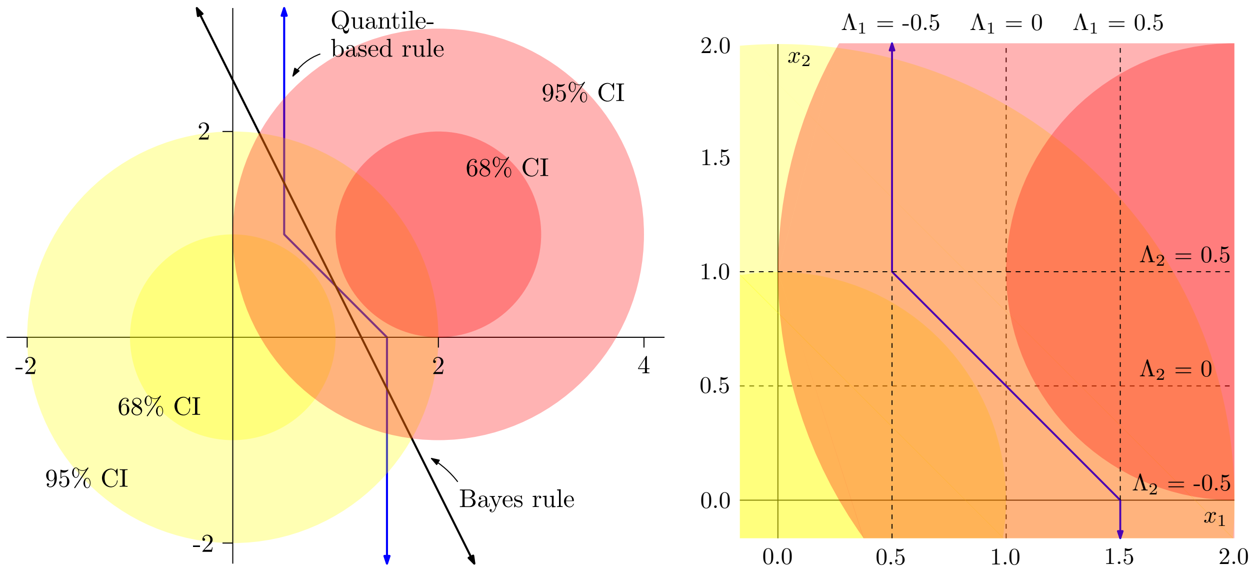
Having described the general form of the composite quantile-based classifiers, we now consider a few illustrative examples of the classifiers in two dimensions. In each of the examples we use the optimal choice of quantile level for each of the features to discriminate between the classes. These are the quantile levels that are chosen by the composite quantile-based classifiers in the limit as the sample size goes to infinity, as a consequence of the convergence results established in Hennig and Viroli (2016). We also set the linear combination coefficients as $(\alpha_0, \alpha_1, \alpha_2) = (0, 1, 1)$ and the prior probability for each class as 1/2.

2.2.5.1 Example 1: Two dimensional Gaussian data

In the first example, shown in Figure 2.2, we consider the classical setting of two Gaussian distributions. The first class, shown in yellow, follows a $N\big((0,0),\, \boldsymbol{I}\big)$ distribution, while the second class, shown in light red, follows a $N\big((2,1),\, \boldsymbol{I}\big)$ distribution. In the left panel, the inner circles for each class represent a contour line for 68% of the mass of the distribution, while the outer circles represent a contour line for 95$\%$ of the mass of the distribution. Additionally the decision rule boundary lines are shown for both the composite quantile-based classifiers method and the Bayes rule. The composite quantile-based classifiers rule has a classification rate of 0.82, while the Bayes rule classification rate is 0.87.

The same contour lines are shown on the right panel, but the figure is zoomed in on the upper-right quadrant of the coordinate system in order to focus on the decision rule boundary in more detail. To begin with, we note that the marginal distributions' Bayes rules have decision rule boundaries with values of 1 and $1/2$ for component 1 (the horizontal axis) and component 2 (the vertical axis), respectively, which corresponds to the median quantile for each component (i.e. $\theta_1 = \theta_2 = 0.5$). It follows that the corresponding population quantiles are given by $F_{01}^{-1}(\theta_1) = 0$, $F_{11}^{-1}(\theta_1) = 2$, $F_{02}^{-1}(\theta_2) = 0$, and $F_{11}^{-1}(\theta_2) = 1$, and furthermore that $\Lambda_1(1, \theta_1) = 0$ and $\Lambda_2(0.5, \theta_2) = 0$. The slope of the line within the box bounded by the within-class quantiles is -1. Once the second component (on the vertical axis) is either no less than 1 or no greater than 0 then Λ_2 becomes a fixed value and the second component becomes a free variable, while the first component is fixed at 0.5 or 1.5, respectively.

Figure 2.2: Composite quantile-based classifier for Gaussian distributions setting.

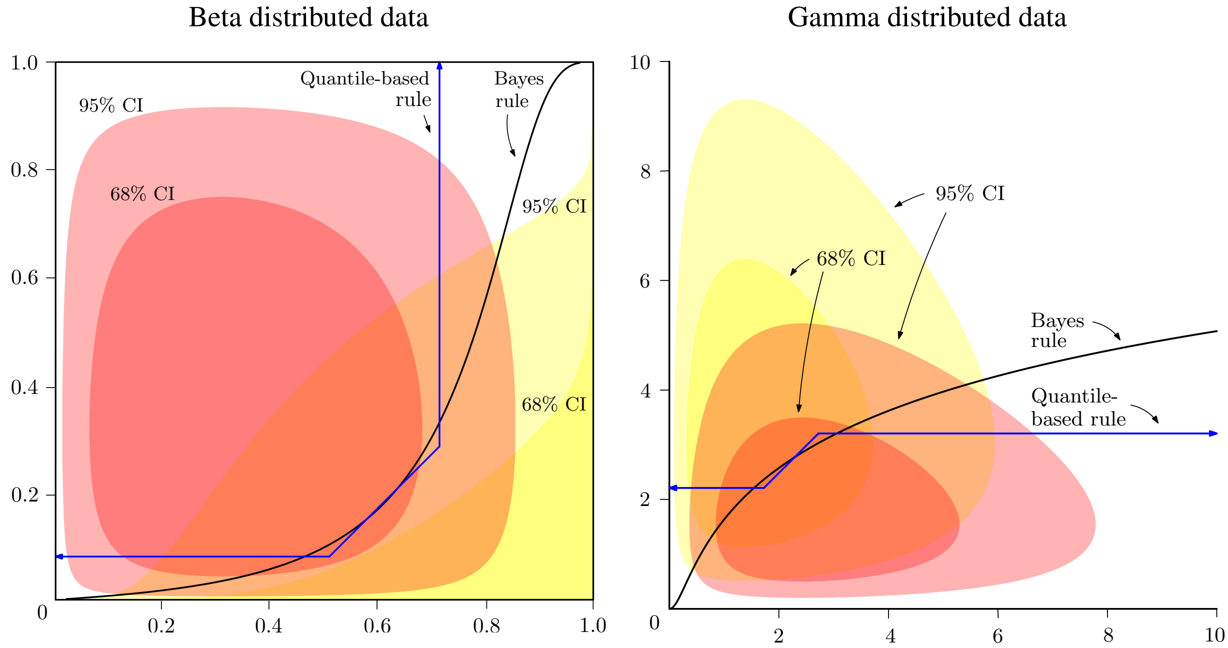


2.2.5.2 Example 2: Two dimensional beta and gamma distributed data

In the second example, shown in Figure 2.3, we consider settings where the shape of the component-wise within-class marginal distributions varies across populations. In the left panel, we consider populations where the within-class distributions follow independent beta distributions, and in the right panel we consider populations where the within-class distributions follow independent gamma distributions. The parameters of the beta and gamma distributions were chosen at random. As in the previous example, contour lines denoting 68% and 95% of the distributional mass for each population are shown. The Bayes rule decision boundary is shown as the black line and the composite quantile-based decision rule boundary is shown as the blue line. The composite quantile-based classifiers rule has a classification rate of 0.84 compared to 0.87 for the Bayes rule for the beta distributed data, and a classification rate of 0.75 compared to 0.76 for the gamma distributed data.

In both examples we see that the Bayes rule decision boundary is nonlinear, so we can only at best hope to approximate it. Fortunately, in both examples the composite quantile-based decision rule boundary follows the Bayes rule decision boundary well in the higher-density regions of the distributions. As the Bayes rule decision boundary moves into low-density distributional regions the approximation is less accurate due to the form of the composite quantile-based classifiers. In our experience this is typical of composite quantile-based classifiers: the approximation of the Bayes rule is often good at the center of the data and is less accurate in the tails.

Figure 2.3: Composite quantile-based classifier for beta- and gamma-distributed data settings.



One limitation of composite quantile-based classifiers that is evident from these examples is that the classification rate does not in general achieve the Bayes rule classification rate even in the limit. This is a consequence of the piecewise linear form of the decision rule boundary. We view this limitation as a trade-off between simplicity of the classifier and its ability to perform well in all situations. A similar property holds for tree-based methods, which perform a vertical or horizontal cut at each stage. Because of the simple nature of the classifier it is well-suited for high-dimensional problems, even in the presence of limited data. For settings where the Bayes rule decision boundary is linear as in Example 1, a variant of composite quantile-based classifiers is discussed in the following section that can achieve the Bayes rule classification rate.

2.3 Numerical results

In this section we compare composite quantile-based classifiers with that of nine other classification methods: quantile-based classifiers Hennig and Viroli (2016), FANS Fan et al. (2016), penalized linear regression (ℓ_1 penalty) Park and Hastie (2007), support vector machine (radial kernel) Cortes and Vapnik (1995), k-nearest neighbor Cover and Hart (1967), naive Bayes Hastie et al. (2009), nearest shrunken centroids

Tibshirani et al. (2002), penalized LDA Witten and Tibshirani (2011), and decision trees Breiman et al. (1984). Detailed descriptions of the settings and software implementations for each of these methods are provided in the supplementary materials.

2.3.1 Simulation examples

We consider three simulated data scenarios in the paper; additional simulation examples and results are provided in the supplementary materials. Within each scenario, we consider combinations of sample sizes $n = 50, 250, 500$ and data dimensions $p = 50, 250, 500$, and in every setting we consider equal numbers of observations for each class. Additionally, the number of discriminatory features remains fixed at 50 throughout all of the simulations, so that when p is larger than 50 then the remaining variables are noise variables.

In the first scenario we consider different distributions for each of 5 blocks of features. In more detail, the data are sampled from a Gaussian distribution with components $(Z_1, \ldots, Z_{50})$, and then the features are evenly split into 5 blocks of 10 and the following transformations performed to each block: (i) $V_{ij} \sim Z_j$ and $W_{ij} \sim Z_j + 0.2$, (ii) $V_{ij} \sim \exp(Z_j)$ and $W_{ij} \sim \exp(Z_j) + 0.2$, (iii) $V_{ij} \sim \log|Z_j|$ and $W_{ij} \sim \log|Z_j| + 0.1$, (iv) $V_{ij} \sim Z_j^2$ and $W_{ij} \sim Z_j^2 + 0.2$, and (v) $V_{ij} \sim |Z_j|^{1/2}$ and $W_{ij} \sim |Z_j|^{1/2} + 0.1$. In Table A.3 we consider uncorrelated data for the underlying Gaussian distribution, and in Table A.4 we consider autoregressive correlation.

In the second scenario, we consider data with features that follow independent gamma distributions with different distributional shapes within each feature. The gamma distribution parameters were sampled as follows. Two shape and scale parameters for each feature were each sampled from a $unif(0.5, 2)$ distribution. Then for each class and feature, each parameter was transformed by taking the absolute value of some additive Gaussian random noise. So for example, suppose that α_j and γ_j are the shape and sclae parameters drawn for the j-th feature. Then for each class the distributional parameters are each sampled from $|\alpha_j + N(0, \sigma_j^2)|$ and $|\gamma_j + N(0, \sigma_j^2)|$ distributions. The parameters $\sigma_1, \ldots, \sigma_{50}$ were given by a fixed sequence each with values between 0.1 and 0.2. Once the shape parameters were sampled, the same parameters were used for every replicate and every simulation study. This is the setting shown in Table A.6.

In the third scenario, we consider data with features that follow independent mixture Gaussian distributions. The first population is a mixture of three Gaussian distribution each with variance 1 and with means -3, 0, and 3, and prior probabilities of 0.2, 0.6, and 0.2 respectively. The second population is a mixture of two Gaussian distributions each with variance 1 and with means -1.5 and 1.5, and prior probabilities of 0.5, and 0.5 respectively. This is the setting shown in Table 2.2b.

The classifier proposed in this paper is abbreviated as CQC. Results for the second variant of composite quantile-based classifiers where the transformed quantile distances data is augmented by the original data are also shown and are abbreviated as CQC augmented. We also consider the multimodal variant of the quantile classifier in the simulation studies, which we abbreviate as CQC multimodal, and the multimodal variant augmented by the original data, which we abbreviate as CQC multimodal augmented.

When the amount of training data are limited, we find that composite quantile-based classifiers can suffer from instability in the choice of quantile levels. To mitigate this behavior, we suggest a modification to the usual composite quantile-based classifiers. Rather than selecting the quantile levels independently for each feature, we instead restrict the quantile level to be constant across the features as is done for quantile-based classifiers. Then, for a fixed choice of quantile level we choose the linear combination coefficients as was proposed for composite quantile-based classifiers; we call this blend of approaches the hybrid approach. Then we choose the classifier with the best classification rate from one of the classifiers based on

$$\alpha_{0k} + \sum_{j=1}^{p} \alpha_{jk} \Lambda_j(z_j,\, k/(q+1)), \qquad k = 1, \ldots, q,$$

for some grid size q. This approach has the advantages of choosing quantile levels that are more stable and doesn't require splitting the data to train the quantile levels, at the cost of losing flexibility in the classifier. In order to choose between the regular and hybrid approach we can choose the classifier with the better classification rate on the cross-validation testing sets. It is noted in the discussion of the simulations whenever the hybrid approach is used.

Simulation results for the block transformed Gaussian setting are shown in Tables A.3 and A.4. In these settings composite quantile-based classifiers and FANS typically perform better than other classifiers. Decision trees also perform well in some settings, which suggests that there are some relatively discriminative features. When the quantity of the training data decreases to $n = 50$ we see a sharp rise in the misclassification rates for all methods. It is in this setting that the hybrid quantile-based classifiers do well. For example, when $n = 50$ and $p = 500$ with uncorrelated data, using the hybrid approach reduces the misclassification rate from 0.351 to 0.215 as compared to composite quantile-based classifiers. We believe the reason that the hybrid approach works well here is that we may have a number of relatively discriminative features that the quantile-based methods approach of selecting quantile levels can effectively key in on, and then the additional linear combination step can help to deal with some of the differences in scale and discriminatory power between the blocks.

Simulation results for the gamma distributed data setting are shown in Table A.6. In this setting, composite quantile-based classifiers and decision trees performed the best, which we see this as a sign of a few features having some degree of separability across the two classes. Since the optimal quantile level varies across all of the features, the additional flexibility of composite quantile-based classifiers to select varying quantile levels results in better performance as compared to quantile-based classifiers.

Simulation results for the mixture Gaussian distributed data setting are shown in Table 2.2b. In this setting the multimodal version of the composite quantile-based classifiers performs the best when the sample size is 250 or 500. When the sample size is small (50 observations) then multimodal composite quantile-based classifiers have some difficulties due to the fact that the observations are split across the observed partitions, thus reducing the sample size for the individual subproblems to too small of a level. Simulation settings show good performance for regular composite quantile-based classifiers for 50 observations in most settings, so we would expect that 50 observations for each subproblem would be sufficient to achieve good performance for the overall multimodal classifier.

2.3.2 Application to spam email classification

For a real data study, we consider a spam email data set with a total of 4,601 observations and 57 features. The attributes are, for example, the percentage of specific words or phrases in the email (e.g. money, free, order), the average and maximum run lengths of uppercase letters, and the total number of uppercase letters. The features for this data set are typically highly skewed: often 90% of the data have no occurrences of a word while a few emails have a high rate. In this data set 39% of the emails in the data set were spam emails, versus 61% non-spam emails.

Various settings were considered for this data as follows. For each setting, a certain amount of data was randomly selected to be the training data, and the misclassification rate was then based on the remaining holdout data. The number of observations used to train the classifiers for the various settings was 100, 250, 500, and 1000 observations, respectively. The misclassification rate as the number of training observations increased for more than 1,000 training observations was nearly constant for all of the classifiers and consequently is not shown.

Table 2.1: Simulation study: misclassification results for block transformed data.

(a) Correlation coefficient = 0 with $p = 50$

	$n = 50$	$n = 250$	$n = 500$
CQC	0.181 (0.03)	**0.043 (0.01)**	**0.029 (0.01)**
CQC multimodal	0.217 (0.06)	0.045 (0.01)	**0.029 (0.01)**
CQC augmented	**0.176 (0.05)**	0.048 (0.01)	0.030 (0.01)
CQC multimodal augmented	0.225 (0.05)	0.046 (0.01)	0.031 (0.01)
Quantile-based classifiers	0.223 (0.03)	0.112 (0.02)	0.085 (0.01)
FANS	0.320 (0.08)	0.068 (0.01)	0.027 (0.01)
FANS2	0.323 (0.06)	0.063 (0.01)	0.028 (0.01)
Penalized logistic regression	0.431 (0.03)	0.322 (0.02)	0.291 (0.02)
Support vector machine	0.386 (0.02)	0.306 (0.02)	0.284 (0.01)
k-nearest neighbor	0.464 (0.02)	0.448 (0.02)	0.431 (0.02)
Naive Bayes	0.452 (0.01)	0.396 (0.02)	0.374 (0.02)
Nearest shrunken centroids	0.422 (0.05)	0.339 (0.01)	0.330 (0.02)
Penalized LDA	0.374 (0.03)	0.301 (0.02)	0.283 (0.01)
Decision trees	0.374 (0.08)	0.078 (0.01)	0.040 (0.01)

(b) Correlation coefficient = 0.8 with $p = 250$

	$n = 50$	$n = 250$	$n = 500$
CQC	**0.228 (0.03)**	**0.115 (0.01)**	0.091 (0.00)
CQC multimodal	0.339 (0.06)	**0.115 (0.01)**	0.094 (0.01)
CQC augmented	0.293 (0.05)	0.118 (0.01)	0.092 (0.00)
CQC multimodal augmented	0.383 (0.06)	0.116 (0.01)	0.091 (0.01)
Quantile-based classifiers	0.263 (0.02)	0.154 (0.02)	0.126 (0.03)
FANS	0.428 (0.08)	0.128 (0.01)	**0.079 (0.01)**
FANS2	0.402 (0.07)	0.139 (0.01)	0.085 (0.01)
Penalized logistic regression	0.477 (0.03)	0.432 (0.03)	0.421 (0.02)
Support vector machine	0.452 (0.02)	0.428 (0.03)	0.416 (0.02)
k-nearest neighbor	0.463 (0.02)	0.450 (0.02)	0.452 (0.03)
Naive Bayes	0.466 (0.02)	0.436 (0.02)	0.435 (0.02)
Nearest shrunken centroids	0.434 (0.02)	0.388 (0.02)	0.392 (0.02)
Penalized LDA	0.460 (0.03)	0.403 (0.02)	0.397 (0.02)
Decision trees	0.443 (0.07)	0.118 (0.02)	0.081 (0.01)

Table 2.2: Simulation study: misclassification results for varying within-class distributional shapes

(a) Gamma distributed features with $p = 50$

	$n = 50$	$n = 250$	$n = 500$
CQC	**0.128 (0.03)**	0.052 (0.01)	0.035 (0.01)
CQC multimodal	0.158 (0.04)	**0.051 (0.01)**	0.037 (0.01)
CQC augmented	0.153 (0.04)	0.054 (0.01)	**0.034 (0.01)**
CQC multimodal augmented	0.191 (0.05)	0.053 (0.01)	**0.034 (0.01)**
Quantile-based classifiers	0.156 (0.01)	0.062 (0.02)	0.045 (0.01)
FANS	0.389 (0.10)	0.107 (0.02)	0.069 (0.01)
FANS2	0.363 (0.09)	0.116 (0.02)	0.073 (0.01)
Penalized logistic regression	0.493 (0.01)	0.500 (0.02)	0.501 (0.02)
Support vector machine	0.492 (0.01)	0.462 (0.02)	0.424 (0.01)
k-nearest neighbor	0.495 (0.01)	0.489 (0.02)	0.486 (0.02)
Naive Bayes	0.424 (0.02)	0.369 (0.02)	0.341 (0.02)
Nearest shrunken centroids	0.506 (0.02)	0.506 (0.02)	0.503 (0.01)
Penalized LDA	0.498 (0.00)	0.506 (0.02)	0.501 (0.01)
Decision trees	0.272 (0.11)	0.060 (0.01)	0.038 (0.01)

(b) Mixture Gaussian distributions with $p = 250$

	$n = 50$	$n = 250$	$n = 500$
CQC	0.464 (0.03)	0.234 (0.01)	0.187 (0.02)
CQC multimodal	0.400 (0.05)	**0.083 (0.01)**	**0.054 (0.01)**
CQC augmented	0.481 (0.02)	0.264 (0.02)	0.195 (0.02)
CQC multimodal augmented	0.442 (0.03)	0.096 (0.01)	0.061 (0.01)
Quantile-based classifiers	**0.359 (0.03)**	0.192 (0.02)	0.158 (0.01)
FANS	0.444 (0.03)	0.117 (0.02)	0.056 (0.01)
FANS2	0.472 (0.06)	0.127 (0.02)	0.057 (0.01)
Penalized logistic regression	0.499 (0.01)	0.495 (0.02)	0.499 (0.02)
Support vector machine	0.504 (0.01)	0.417 (0.05)	0.374 (0.02)
k-nearest neighbor	0.488 (0.01)	0.481 (0.01)	0.487 (0.01)
Naive Bayes	0.431 (0.02)	0.318 (0.03)	0.270 (0.02)
Nearest shrunken centroids	0.502 (0.01)	0.492 (0.02)	0.496 (0.01)
Penalized LDA	0.506 (0.01)	0.497 (0.02)	0.506 (0.02)
Decision trees	0.491 (0.02)	0.398 (0.06)	0.334 (0.04)

We observe that augmented quantile-based classifiers have nearly the lowest misclassification rate across all of the settings. Penalized logistic regression is one of the better competitors which suggests that a linear decision rule boundary is a reasonable choice of boundary for this problem. The augmented FANS classifier is another strong competitor which makes sense given that it also uses the original features, and has a close relationship to composite quantile-based classifiers.

Table 2.3: Misclassification rates for the spam email data set for varying numbers of observations used as training data.

	Number of observations used to train			
	$n = 100$	$n = 250$	$n = 500$	$n = 1000$
CQC	0.130 (0.04)	**0.088 (0.01)**	0.080 (0.01)	0.068 (0.01)
CQC augmented	**0.123 (0.03)**	0.090 (0.01)	**0.079 (0.01)**	**0.064 (0.00)**
Quantile-based classifiers	0.319 (0.03)	0.307 (0.01)	0.305 (0.01)	0.302 (0.01)
FANS	0.153 (0.03)	0.101 (0.01)	0.095 (0.01)	0.079 (0.00)
FANS2	0.146 (0.02)	0.102 (0.01)	0.089 (0.01)	0.074 (0.00)
Penalized logistic regression	0.140 (0.03)	0.116 (0.01)	0.100 (0.01)	0.081 (0.01)
Support vector machine	0.249 (0.11)	0.130 (0.05)	0.100 (0.01)	0.081 (0.00)
k-nearest neighbor	0.315 (0.01)	0.299 (0.01)	0.282 (0.01)	0.251 (0.01)
Naive Bayes	0.391 (0.03)	0.352 (0.02)	0.314 (0.02)	0.295 (0.02)
Nearest shrunken centroids	0.148 (0.03)	0.152 (0.02)	0.153 (0.02)	0.134 (0.01)
Penalized LDA	0.138 (0.02)	0.129 (0.01)	0.121 (0.01)	0.113 (0.01)
Decision trees	0.204 (0.03)	0.156 (0.02)	0.133 (0.02)	0.108 (0.01)

2.4 Discussion

In this paper we propose a family of classification rules based on the marginal quantiles of the class features. The univariate quantile classifier is equal to the Bayes rule for the optimal choice of quantile level and under some assumptions. Motivated by this, we consider univariate quantile classifiers as a starting point for constructing a multivariate classifier. The multivariate classifiers considered in this paper are constructed in a two step process. First, the marginal quantiles of the data are estimated and an estimate of the optimal quantile level for each component is calculated. Secondly, a rule for combining the information provided by an observation's quantile distances to the marginal within-class quantile is constructed through a linear combination obtained using penalized linear regression. We observe competitive performance of composite quantile-based classifiers in simulation examples and a spam email classification application. The composite quantile-based classifiers decision rule is computationally efficient to train and the decision rule boundary has a simple piecewise-linear form that is well-suited for high-dimensional settings.

Proofs to the theorems in this paper are presented in the supplementary materials.

CHAPTER 3: BIOACTIVE PEPTIDE DISCOVERY

Difficult-to-treat fungal and bacterial infections are increasingly commonplace, new viral diseases are emerging and spreading rapidly, and cancer remains a leading cause of death worldwide (Klevens et al., 2007). In the United States, there are almost two million hospital-acquired bacterial infections each year, resulting in over 100,000 deaths. The emergence of ESKAPE (Enterococcus faecium, Staphylococcus aureus, Klebsiella pneumoniae, Acinetobacter baumannii, Pseudomonas aeruginosa, and Enterobacter species) pathogens has led to a significant increase in multidrug-resistant infections in the clinic with associated increases in morbidity and mortality (Boucher et al., 2009). In addition, the recent emergence, resurgence, and spread of viruses, including Zika, SARS, West Nile, Ebola, and MERS (Fletcher and Moreno, 2012), for which there are limited or no direct-acting antivirals, highlight the susceptibility of the human population to future potentially untreatable pandemics. Furthermore, despite continued progress in anticancer therapeutics, limitations to current lead compounds of nonspecific toxicity, poor drug penetration, and multidrug resistance have emphasized the need for discovery of anticancer therapeutics with novel mechanisms of action. With many bacteria now unresponsive to multiple classes of antimicrobial compounds and cancer being the leading cause of death worldwide, there is an undeniable and desperate need to identify novel pharmacological chemistries and accelerate their development through new methods and innovative technologies.

Natural products have long been sources of virtually all traditional medicinal preparations and have been the single most successful source of lead compounds for drug discovery (Harvey et al., 2015). Specifically, plants have played a significant role in the treatment of human ailments since prehistoric times. The teas and tinctures of times past are one source of drug discovery, allowing the ethnobotanically guided isolation and characterization of pharmacologically active compounds for the treatment of bacterial and fungal infections, cancers, and other ailments. Despite historical relevance and past success, challenges associated with natural product discovery have slowed progress in drug discovery efforts. In addition, natural product discovery efforts have largely focused on small molecule constituents. However, recent discoveries have revealed ribosomally synthesized, post-translationally modified peptide natural products (RiPPs) with substantial structural diversity and bioactivity potential, including novel mechanisms of action (Arnison et al., 2013;

Essig et al., 2014). While traditionally studied for antifungal and antibacterial properties, recent studies have piqued interest in these peptides as potential anticancer therapeutics (Schweizer, 2009). Peptides offer several advantages over other small and large molecule drug candidates – including greater efficacy, selectivity, and specificity, relative to small organic molecules, better tissue penetration, and reduced immunogenicity and manufacturing costs, relative to proteins/antibodies. Advances in peptide modification, formulation, and delivery methods can address known limitations of peptidic drug candidates (Vlieghe et al., 2010), including modification of peptide length/content to increase selectivity (Mandal et al., 2014), stapling, and/or peptidomimetic conversion techniques to improve pharmacokinetic properties (Walensky and Bird, 2014), as well as encapsulation methods to protect from proteolytic degradation. However, RiPPs fall outside the scope of standard therapeutic screening approaches and elude detection via standard -omic workflows, because of their large size, structural diversity, and high level of post-translational modification; therefore, systematic approaches for discovery and characterization are in nascent development (Mohimani et al., 2014). Developing natural product screens for antimicrobially active RiPPs has the potential for discovery of new bioactive compounds with novel mechanisms of action able to address the growing problem of antimicrobial resistance.

Current methods for RiPP discovery often rely on bioassay-guided fractionation (Henriques et al., 2012) or genomic mapping (e.g., Pep2 Path, RiPPquest) to facilitate downstream analysis (Medema et al., 2014; Skinnider et al., 2016). Relying on iterative rounds of chromatographic separations, bioassay- guided fractionation is extremely time-consuming and provides no structural information until late in the discovery process. In addition, this approach often leads to replication of previously known compounds, as a bias toward highly abundant and highly active compounds is evident. While alternative genomic mapping approaches can provide structural information from the beginning, prior genomic sequencing and knowledge of peptide biosynthetic pathways is required. Furthermore, this approach is unable to provide direct bioactivity information on the target peptide, thus necessitating downstream isolation and activity screening to determine function and biological activity. Recent approaches to address some of these limitations have been proposed (e.g., compound activity mapping); however, these platforms are specific to small molecule identification and do not translate efficiently toward advancing untargeted proteomics approaches (Kurita et al., 2015).

The research presented in this chapter is the result of a collaboration in Kirkpatrick et al. (2017), with the resulting methodology named the PepSAVI-MS pipeline. The goal of the research is to develop a bioinformatics and statistical pipeline for the screening and identification of bioactive peptides from natural product sources. For streamlined identification of bioactive peptides, the platform relies on mass spectrometry for data collection with subsequent data processing and statistical analysis to assign bioactivity to individual components. Proof-of-principle studies for a known antimicrobially active RiPP from the plant Viola odorata are presented to validate this workflow. In addition, we use this pipeline to explore novel antibacterial and anticancer applications of the Viola odorata cyclotide, cycloviolacin O2.

3.1 Data generation

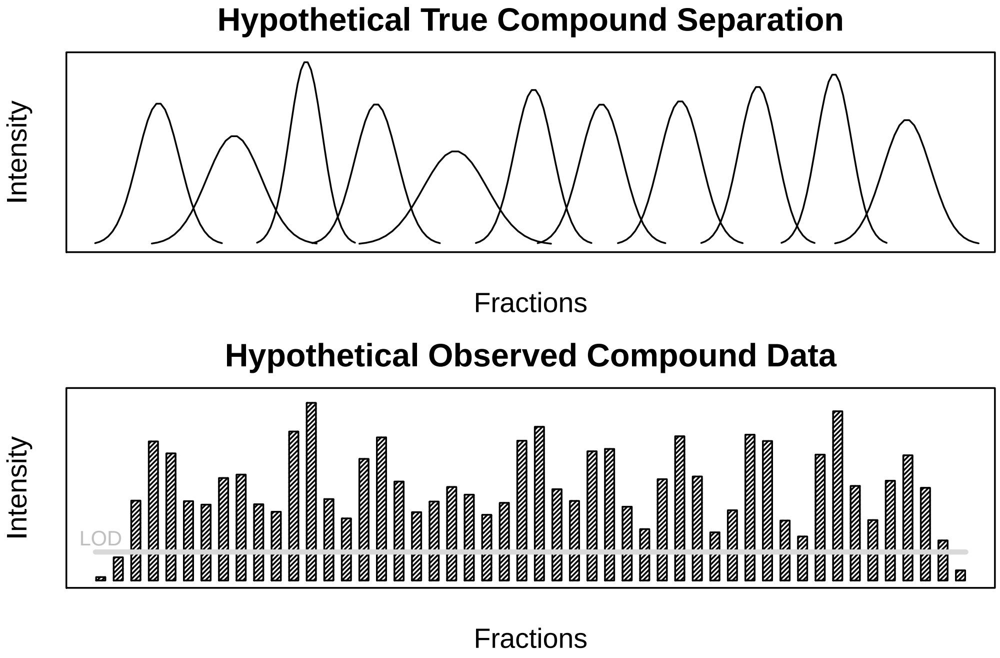
3.1.1 Explanatory data generation

Viola odorata seedlings were grown to mature rosette stage under standard greenhouse conditions. Seedlings were planted in nutrient-rich soil under controlled temperature and light cycle conditions until aerial tissue was harvested using flash freezing under liquid nitrogen. Frozen tissue was ground using a mortar and pestle and aqueous solution was created by introducting acetic acid with protease inhibitors. The sample was then concentrated using vacuum centrifugation to generate the final crude extract.

The quantity of each of the candiate compounds in the extract can be measured using mass spectrometry. However, in order to distinguish compounds with similar spectrometry readings and additionally to have greater statistical capability to detect compounds with antimicrobial effects, the eluent generated from the viola odorata tissue is subjected to liquid chromatography separation. This process involves forcing the eluent through a a chromatographic column. The analates flow through the column at varying speeds depending on their chemical composition, and every few minutes a new vial collects the eluate. As a function of time, the quantity that a given analate is observed follows a unimodal curve. See Figure 3.1 for a hypothetical representation for how a sequence of analates with similar spectrometry readings can be observed in the spectrometer over time. In the top panel, we show a hypothetical true quantity for a given mass spectometer range and where each curve represents a unique compound. In the bottom panel we show how quantities in the top panel are observed at discrete time slices across the fractions.

The mass spectrometer measures the quantity of a compound at a given mass to charge ratio (m/z) and charge value. Data was collected for a total of 33 fractions and 10,902 mass to charge ratios. However, a given compound may span multiple m/z levels, so one of our tasks in the upcoming data processing section is to combine readings for different m/z levels belonging to the same compound.

Figure 3.1: Liquid Chromotography Representation



3.1.2 Outcome data generation

The outcome variables are a measure of the quantity of growth inhibition in a given pathogen or cancer cell line after introducing a sample of the aqueous solution generated from the viola odorata tissue into a petri dish with the pathogen or cancer cell line. Peptide libraries were assayed for growth inhibition against the following pathogens: *E. coli*, *E. faecium*, *S. aureus*, *K. pneumoniae*, *A. baumannii*, *P. aeruginosa*, *E. cloacae*, *F. graminearum*, and cancer cell lines: breast cancer, prostate cancer, and ovarian cancer. Library fractions were incubated with a microbial or cancer cell culture in a manner such that the presence of bioactive peptides in a given fraction will result in inhibition of culture growth during the incubation period. For bacterial assays, the remaining viable cells were quantified indirectly by spectrophotometric measurement of the irreversible intracellular bioreduction of resazurin. For anticancer bioactivity, cytotoxicity assays were performed using MTT-based assays to measure mitochondrial succinate dehydrogenase activity with absorbance measurements at 570 nm. Values for each fraction are compared to positive and negative controls containing a known therapeutic or water, respectively, to determine a percent activity of each library fraction, where a small value of remaining viable cells indicates high activity. Percent activity of each well was calculated using the formula:

$$\text{percent activity} = \left(1 - \frac{\text{Response of fraction} - \text{Response of positive control}}{\text{Response of negative control} - \text{Response of positive control}}\right) \times 100 \quad (3.1)$$

where response refers to relative fluorescence units for antibacterial assays and absorbance for anticancer assays. There were a total of either 3 or 4 measurements conducted for each of the pathogens or cancer cell lines for each of the fractions.

3.2 Processing and filtering the explanetory data

3.2.1 Recovering the candidate compounds

The preprocessing and binning procedure contains a number of criteria to exclude unwanted compounds and thereby restrict compounds to those of potential interest. Retention time limitations are put in place to eliminate background ions that are detected either before (equilibration) or after (wash) the gradient is applied; for the gradient performed in our laboratory, 14 and 45 minutes are appropriate retention time lower and upper bounds. Mass and charge limitations of 2 to 15 kDa and +2 to +10, respectively, were chosen to restrict compounds to the mass and charge ranges of known bioactive peptides.

Mass spectrometry features that satisfy this initial exclusion process are then consolidated when features are believed to belong to the same underlying compound. This consolidation step requires choices to be made through the specification of criteria deeming two MS features to be considered the same compound. An m/z difference of 0.05 Da was chosen for our data analysis to be the maximum difference for which two compounds could have in m/z to be considered to belong to the same compound. The value of 0.05 Da was chosen by performing multiple LC-MS runs studying the variation in the observations; this value was chosen so that fluctuations in mass accuracy and retention time for the test data allowed the same peak to be picked multiple times despite representing the same compound.

The other criterion that must be met for two MS feature observations to be considered to belong to the same compound is that the retention times for the two features cannot be more than a specified amount of time apart. For this analysis, because we used prior retention time alignment, we chose to allow all MS features that were below the threshold for m/z difference and had the same charge state to be considered to belong to the same compound.

The precise algorithm used to filter and combine mass spectrometry observations that are believed to belong to the same underlying compound is described as follows. In the first step, all observations must satisfy each of the following criteria for inclusion in the binning process.

1. Each observation must have its peak elution time occur during the fractions of interest.
2. Each observation must have a mass that exceeds the spectrometer's limit of detection.
3. Each observation must have an electrical charge state that falls within the expected range of known bioactive peptides.

Once that a set of observations satisfying the above criteria is obtained, then a second step attempts to combine observations believed to belong to the same underlying compound. The algorithm considers two observations that satisfy each of the following criteria to belong to the same compound.

1. The absolute difference in Daltons of the mass-to-charge value between the two observations is less than the prespecified value.
2. The absolute difference of the peak elution time between the two observations is less than the prespecified amount.
3. The electrical charge state must be the same for the two observations.

Then the binning algorithm is defined as follows. Consider an observation that satisfies the inclusion criteria; this observation is compaired pairwise with every other observation that satisfies the inclusion criteria. If a pair of observations satisfies the criteria determining them to belong to the same underlying compound then the two observations are merged into a single observation. The two previous compounds are removed from the working set, and the process starts over with the newly created observation. The process repeats until no other observation in the working set meets the criteria determining it to belong to the same underlying compound as that of the current observation; at this point it is considered that all observations belonging to the compound have been found, and the process starts over with a new observation.

The merging process has not yet been defined; it is performed by averaging the mass-to-charge values and peak elution times, and summing the mass spectrometry intensities at each fraction. Although observations are merged pairwise, when multiple observations are combined in a sequence of pairings, the averages are given equal weight for all of the observations. In other words, if a pair of observations are merged, and then a third observation is merged with the new observation created by combining the original two, then the mass-to-charge value and peak elution time values of the new observation are obtained by summing the values for each of the three original observations and dividing by three. The merging process for more than three observations is conducted similarly.

Having described the binning algorithm, it is apparent that there are scenarios in which the order in which observations are merged affects the outcome of the algorithm. Since it seems that a minumum requirement of any binning algorithm is that the algorithm is invariant to the ordering of the observations in the data, this algorithm abides by the following rules. The observations in the data are sorted in increasing order by mass-to-charge value, peak elution time, and electical charge state, respectively. Then when choosing an observation to compare to the rest of the set, we start with the observation at the top of the sort ordering, and compare it one-at-a-time to the other elements in the set according to the same ordering. When a consolidated observation is complete in that no other observation left in the working set satisfies the merging criteria, then this consolidated observation can be removed from consideration for all future merges.

At the end of the preprocessing and binning procedure, the number of candidate compounds was reduced from 10,902 to 6,258.

3.2.2 Filtering candidate compounds

The next step in the pipeline is to remove any potential candidate compounds with observed abundances for which it is unlikely that they might be a compound with an effect on the observed bioactivity.

3.2.2.1 Selecting the fraction region of interest

Bioactivity regions of interest are selected based on each individual bioactivity data set. Because of the crude nature of SCX, a given peptide should elute over a minimum of three fractions in a Gaussian manner. Because mass spectrometry is more sensitive than the bioactivity assays, the defined bioactivity region can extend 1-2 fractions beyond the visible activity range on either end. The bioactivity regions for each species that were deemed to be active are depicted by blue bars in Figure 3.2. The regions chosen for each bioactivity data ranged from a size of as small as 3 fractions to as large as 6 fractions, and each region was contained within the window of fractions 17-25.

For typical analyses we recommend filtering the MS dataset using the region chosen as dictated by the bioactivity data (as described above). However, for the data analysis, in the interest of simplicity of presentation and since a single region rather tightly encapsulates the chosen region for all of the bioactivity datasets, we simply filtered the MS dataset with the region chosen to be the smallest one that included the chosen region for every individual datset, namely fractions 17-25.

3.2.2.2 Filtering criteria selection

In addition to selecting the region of interest for which to filter by, we must also specify a bordering region; this is the region that we consider when filtering for candidate compounds by criterion 3: a candidate compound's elution in the bordering region cannot be greater than a specified proportion of its maximum abundance, and criterion 4: a candidate compound must have a nonzero level of elution in the right adjacent fraction to the fraction with the maximum elution. For this data analysis, we specified the most conservative choice for bordering region, which is that every fraction not in the region of interest is to be considered the bordering region.

Another criterion that we filter by is that a candidate compound may not by more abundant in the bordering region than some specified proportion of its maximum abundance in the region of interest. Allowing for a small nonzero elution outside of the region of interest accounts for small fluctations from absolute 0 that are most likely due to noise in the instrumentation readings. The allowable proportion of abundance relative to the maximum is data dependent; we selected a value of 0.01 by inspecting the distribution of MS features in the data that fit the profile of what we might expect for a compound with an effect on bioactivity levels in the region.

Another filtering criterion that we consider for a candidate compound is a minimum value for the maximum intensity value over the LC-MS profile. This level should be chosen to reflect the appropriate amount of noise apparent in the instrumentation; for our analysis we selected 1000 Da as an appropriate value.

Finally, an option to filter the data by a maximum allowed charge state is included in the pipeline. This has already been considered in the consolidation step, but is included in the filtering step in the event that other researchers do not wish to use the consolidation function with the data analysis in hand. We selected a maximum charge state of +10, consistent with our choice in the binning step.

After this filtering of the data, this reduced the number of candidate compounds from 6,258 to 225.

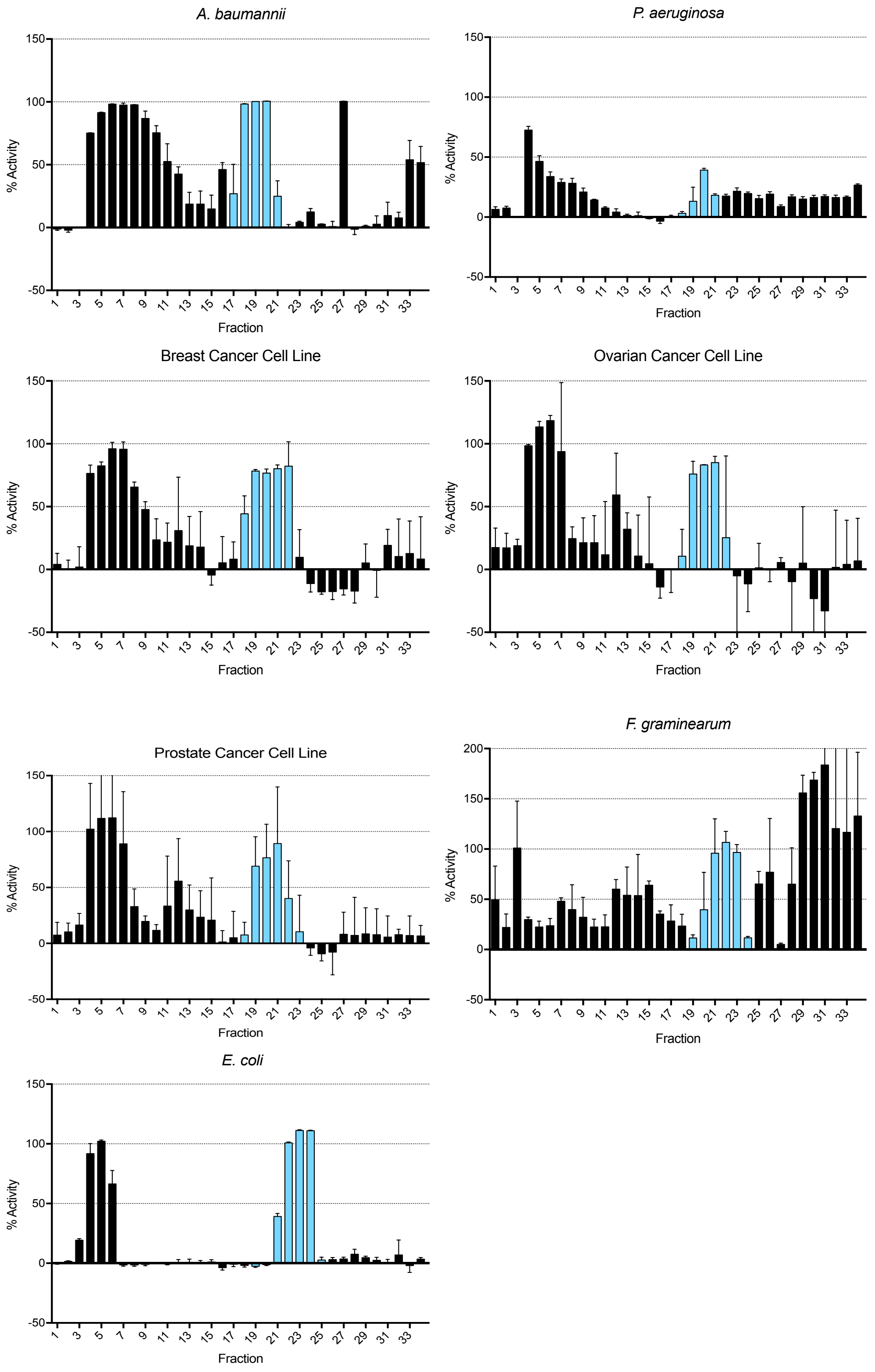


Figure 3.2: *Viola odorata* peptide library bioactivity against all pathogens in which strong activity was seen in the region of cyO2 elution. Average percentage of activity values for each fraction are plotted with error bars representing +1 standard deviation. The fractions selected as the region of interest are shown using blue bars.

3.3 Candidate compound ranking and validation

In this section potential candidate compounds are ranked with the goal of facilitating the investigation of compounds with a deleterious effect on each of the bioactivity peptide libraries. Not that for each library the mass spectrometry data remains the same, and additionally that the region of interest changes over the peptitde libraries according to the selections shown in Figure 3.2.

3.3.1 Candidate compound ranking procedure

Constructing a ranking procedure for the compounds was highly constrained due to the availability of the data. Despite our best efforts, the number of candidate compounds was reduced to 225, while we had available only between 12 and 20 observed outcomes for a given pathogen or cancer cell line. Due to this dearth of data, our procedure is constructed to require only a few observations to produce reasonable results.

The ranking procedure works by selecting a choice of quadratic penalty parameter, and for fixed value of quadratic penalty parameter tracking the order in which the coefficients corresponding to candidate compounds first become nonzero along the elastic net (Zou and Hastie, 2005) path as the ℓ_1 penalty parameter changes. Then it is presumed that compounds with corresponding coefficients that become nonzero earlier in the path may be better candidates for having an effect on bioactivity levels then those with corresponding coefficients that become nonzero later in the path. To select a quadratic penalty parameter we recommend a small nonzero value with the thought that being near to the lasso penalty (Tibshirani, 1996) we might achieve good behavior in terms of variable selection. The reason for not choosing a penalty of 0 is that in that case the elastic net reduces to the lasso model, which can have no more than one less than the number of bioactivity data points of nonzero coefficients, and thus severly reduces the size of the list of candidate compounds generated by this approach. We also note that we do not consider this choice to be data-driven; for our analysis we chose a value of 0.001.

3.3.2 Pipeline validation using cyO2

To validate this pipeline, we demonstrate successful detection and identification of a known AMP from the botanical species *Viola odorata*. *Viola odorata*, commonly known as sweet violet, contains many cyclotides - including cycloviolacin O2 (cyO2). CyO2 is a small, cysteine rich cyclotide comprised of 30 amino acids (MWmonoisotopic: 3138.37 Da), which has been shown to have diverse activity against many

Gram-negative bacteria (*E. coli*, *K. pneumoniae*, and *P. aeruginosa*), as well as several cancer cell lines. Here, we demonstrate successful detection, prioritization, and identification of cyO2 as a means of validating the use of this statistical analysis approach for bioactive peptide discovery.

Each of the rows in Table 3.1 contains a ranking of the cyO@ compound obtained using the procedure for a given outcome variable. The exact m/z and charge values for cyO2 were obtained by comparing the candidate compounds list after the consolidation and filtering process to the known range of values; the m/z and charge pairs corresponding to cyO2 were determined to be (1047.4898, 3), and (1570.2414, 2) in our data. When there is only one ranking for a given pathogen or cancer cell line then this indicates that only one of the two forms of cyO2 survived the screening process.

From the table we see that for the *E. coli*, and breast cancer data sets, at least one charge state corresponding to cyO2 is identified in the first 20 candidate compounds. For the *A. baumannii*, *P. aeruginosa*, and prostate cancer data cets, CyO2 is identified in the first 50 compounds; and for the *F. graminearum* and ovarian cancer data sets cyO2 is identified in the first 100 compounds.

Outcome	Rankings
E. coli	20, 97
A. baumannii	23
P. aeruginosa	41, 97
F. graminearum	97, 112
Breast cancer cells	3
Prostate cancer cells	45
Ovarian cancer cells	99

Table 3.1: Cycloviolacin O2 rankings for each of the outcome variables

3.4 Discussion

It has been previously demonstrated that cyO2 was active against a subset of the ESKAPE pathogens, including K. pneumoniae (Pränting et al., 2010). Results from our study are in agreement with the activity of cyO2 in the cyclotide fractions against E. coli and P. aeruginosa. The MIC for the ESKAPE pathogen representing novel activity, A. baumannii, was determined to be 15 μM, using isolated cyO2, and thus supports that cyO2 is the main contributor to the activity seen in these fractions.

Screening of V. odorata fractions against a panel of human cancer cell lines similarly demonstrates a proof-of-principle, as well as an additional novel finding. In previous studies, cyO2 has been shown to have anticancer activity against multiple breast and ovarian cancer cell lines (Gerlach et al., 2010). Although our screen used different cell lines, strong activity was observed against the breast cancer cell line, as well as the ovarian cancer cell line across fractions in which cyO2 was eluted. In addition, this is the first demonstration of the ability of purified cyO2 to kill PC3 prostate cancer cells. This study not only highlight the wide applicability of this platform but also the capabilities of antimicrobial peptides as broad-spectrum therapeutics.

CHAPTER 4: DAY-SPECIFIC PROBABILITIES OF CONCEPTION

4.1 Introduction

Understanding the predictors of pregnancy is a fundamental goal of human fertility research. Identifying the factors that are protective or promotive of conception can help clinicians and patients to understand the causes of infertility and aid couples who are trying to conceive. However, modeling the biological system can be challenging since the pregnancy outcome is observed at a cycle-specific time scale, while some predictors of pregnancy such as cervical mucus type or hormone levels are associated with pregnancy at a finer time scale. The day-specific probabilities of conception model proposed in Dunson and Stanford (2005) offers a rich expression of this system, however our experience working with the model has indicated to us that there are several areas in which the posterior estimation can be improved. In this paper we propose several modifications to the day-specific probabilities of conception model that provide for more efficient estimation of the predictors of pregnancy.

4.1.1 The biology of human fertilization

Fertilization is defined as the fusion of a human oocyte (egg) and sperm. In the natural fertility context, this occurs sometime after sexual intercourse between a man and a woman. In more detail, during the reproductive portion of her life, the female body undergoes a cyclical process known as the menstrual cycle that prepares the body for a potential fertilization. During ovulation, a single oocyte is released into the Fallopian tube, where it has the opportunity to be fertilized by a sperm cell. The life span of the egg after release is estimated to be about 12-24 hours, during which fertilization must occur for a successful conception to take place during a given menstrual cycle. If fertilization, embryo development, and subsequent implantation do not occur, a woman will menstruate, which starts another cycle and chance to become pregnant.

On the male side of the reproductive process, sperm leave the male's body and enter the female's vagina after vaginal intercourse and ejaculation. They rapidly move into the cervix, and from there, some sperm will travel into the Fallopian tubes. Fertilization of the egg will occur in the distal portion of the fallopian tube. Sperm are viable for up to 5 days, and can reach the egg in as little time as an hour or as long as several days.

4.1.1.1 The fertile window

Given that an egg is viable for a day and sperm are viable and can fertilize an egg for up to 5 days, conception can occur when vaginal intercourse occurs during a window including the 5 days prior to ovulation and the day of ovulation. This time period is known as the *fertile window* in the fertility studies domain.

Furthermore, it is apparent that sexual intercourse status and other factors are irrelevant outside of the fertile window. To simplify modeling of the biological process, it commonly assumed that the size of the fertile window is a constant number of days across all subjects and menstrual cycles.

One complication of this issue is that we don't usually know when ovulation actually occurs. Ovulation can be estimated by using ovulation predictor kits or by the calendar method, which involves counting backwards from the onset of menstrual bleeding. No ovulation marker is completely accurate, so to account for possible data misspecification, the fertile window is commonly widened to 5-11 days so as to ensure that the true fertile window lies somewhere in the extended window. In this paper we proceed under the assumption that the data is correct, although an interesting direction of research would be to explicitly account for the possible shift between the data and the true fertile window in the model.

4.1.1.2 Data collection

Many potential predictors of pregnancy such as age and gravidity status at the time of entering a study can be thought of as fixed quantities. However some predictors of pregnancy such as cervical mucus type change over the course of a menstrual cycle. Increasing the number of times such variables are measured adds burden to study participants and yields diminishing amounts of information. Therefore a good balance regarding the frequency of collected observations must be reached, and a typical time granularity is one entry per day.

Of course the most critical time-varying event affecting pregnancy is the occurrence of sexual intercourse status. Sexual intercourse status is typically recorded as a binary variable delineating either none or no less than 1 occurrence of sexual intercourse due primarily to modeling convenience.

Data is often recorded in-home by study participants, with the data kept in some form of a journal, which may be some sort of physical medium, or may be entered in some form of digital application such as a Web interface.

4.1.2 Time to Conceive Study

In this paper we use our method to analyze the Time to Conceive study, a prospective study conducted from 2008-2010 that observed English-speaking women between 30 and 44 years of age who were attempting to conceive (Evans-Hoeker et al., 2013; Steiner et al., 2014; Crawford et al., 2016; Steiner et al., 2017). Women who had been trying to conceive for more than three months prior to entering the study, or women with a history of infertility or other medical conditions associated with infertility were excluded from the study.

After participants were entered into the study, women completed a questionnaire which queried women for information on various topics such as demographics, medical, surgical, obstetric, gynecologic, and menstrual history, behaviors, partner demographics, height and weight, and pregnancy history. Then, over the course of the study, participants recorded a study diary, which included information on vaginal bleeding, cervical mucus score, pregnancy test results, and vaginal lubricant use.

4.1.3 Upcoming sections

In Section 4.2, we review the existing day-specific probabilities model proposed in Dunson and Stanford (2005). Then in Section 4.3, we propose a new prior specification for the coefficients corresponding to the fertile window that encourages information sharing and consistency with a biologically plausible form. In Section 4.4, we propose a new approach for incorporating missing data into the model for both sexual intercourse status and model covariates. Then in Section 4.5, we perform simulation studies demonstrating the increased efficiency of the prior specifications for the regression coefficients corresponding to the fertile window days, as well as the increased efficiency from incorporating the missing data into the model. Furthermore, in 4.6, we present a real data analysis using the new approaches for the Time to Conceive dataset. Finally, in Section 4.7, we summarize the proposed model for modeling fertility data.

4.2 The day-specific probabilities model

The day-specific probabilities of conception model was introduced in Dunson and Stanford (2005). For completeness, the notation and definitions are reviewed in the following sections.

4.2.1 Model specification

We wish to model the probability of a woman becoming pregnant for a given menstrual cycle as a function of her covariate status across the days of the cycle. Consider a study cohort and let us index

woman i, $i = 1, \dots, n$,

cycle j, $j = 1, \dots, n_i$,

day k, $k = 1, \dots, K$,

where day k refers to the k^{th} day out of a total of K days in the fertile window. Let us write day i, j, k as a shorthand for individual i, cycle j, and day k and similarly for cycle j, k. Then define

Y_{ij}, an indicator of conception for woman i, cycle j,

V_{ijk}, an indicator of conception for woman i, cycle j, day k,

X_{ijk}, an indicator of intercourse for woman i, cycle j, day k.

Then writing $\boldsymbol{X}_{ij} = \big(X_{ij1}, \dots, X_{ijK}\big)$, it can be shown that

$$\begin{aligned}
&\mathbb{P}\left(Y_{ij} = 1 \,\middle|\, \boldsymbol{X}_{ij},\ Y_{i1} = 0, \dots, Y_{i,j-1} = 0\right) \\
&\quad = 1 - \prod_{k=1}^{K} \left\{1 - \mathbb{P}\left(V_{ijk} = 1 \,\middle|\, Y_{i1} = 0, \dots, Y_{i,j-1} = 0,\ V_{ij1} = 0, \dots, V_{i,k-1} = 0\right)\right\}^{X_{ijk}}.
\end{aligned}$$

With this result in mind, we now consider the Dunson and Stanford day-specific probabilities model. Using the same indexing scheme as above, define

$\boldsymbol{u}_{ijk}$, a covariate vector of length q for woman i, cycle j, day k,

$\boldsymbol{\beta}$, a vector of length q of regression coefficients,

ξ_i, woman-specific random effect.

Then writing $\boldsymbol{U}_{ij} = \left(\boldsymbol{u}'_{ijk}, \ldots, \boldsymbol{u}'_{ijk}\right)'$, Dunson and Stanford propose the model:

$$\begin{aligned}
\mathbb{P}\left(Y_{ij} = 1 \,\middle|\, \xi_i, \boldsymbol{X}_{ij}, \boldsymbol{U}_{ij}\right) = 1 - \prod_{k=1}^{K} (1 - \lambda_{ijk})^{X_{ijk}}, \\
\lambda_{ijk} = 1 - \exp\left\{-\xi_i \exp\left(\boldsymbol{u}'_{ijk}\boldsymbol{\beta}\right)\right\}, \\
\xi_i \sim \text{Gamma}\,(\phi, \phi).
\end{aligned}$$

From our previous derivation, we see that we may interpret λ_{ijk} as the day-specific probability of conception in cycle j from couple i given that conception has not already occured, or in the language of Dunson and Stanford, given intercourse only on day k.

Delving further, we see that λ_{ijk} is strictly increasing in $u_{ijkh}\,\beta_h$, where we are denoting u_{ijkh} to be the h^{th} term in $\boldsymbol{u}_{ijk}$ and similarly for β_h. When $\beta_h = 0$ then the h^{th} covariate has no effect on the day-specific probability of conception.

λ_{ijk} is also strictly increasing in ξ_i which may be interpreted as a woman-specific random effect. Also note that specifying the distribution of the ξ_i with a common parameters prevents nonidentifiability between $\mathbb{E}\left[\xi_i\right]$ and the day-specific parameters. Since $\text{Var}\left[\xi_i\right] = 1/\phi$ it follows that ϕ may be interpreted as a measure of variability across women.

4.2.1.1 Auxiliary model specification

In order to make posterior computation more tractable, Dunson and Stanford proposed the following auxiliary data model.

$$\begin{aligned}
&Y_{ij} = \mathbb{1}\left(\sum_{k=1}^{K} X_{ijk} Z_{ijk} > 0\right), \\
&Z_{ijk} \sim \text{Poisson}\left(\xi_i \exp\left(\boldsymbol{u}'_{ijk}\boldsymbol{\beta}\right)\right), \qquad k = 1, \ldots, K.
\end{aligned}$$

Let us further define $W_{ijk} = X_{ijk} Z_{ijk}$ for all i, j, k.

4.3 Estimating fertile window day regression terms

Based on the biology of human fertility Speroff and Fritz (2005) and the existing fertility studies literature Barrett and Marshall (1969); Schwartz et al. (1980); Royston (1982); Wilcox et al. (1995); Colombo and Masarotto (2000); Dunson et al. (2001); Stanford et al. (2003); Lynch et al. (2006), we expect that the effect of the timing of intercourse relative to the ovulation day will follow a unimodal shape. Thus, by incorporating this information into the model, we anticipate that this will result in more accurate posterior estimation.

4.3.1 Prior specification of the fertile window regression coefficients

Let $S(d)$ be a decay function such that $S(-\infty) = S(\infty) = 0$ and $S(0) = 1$. Further let m index the day associated with peak fertility. Then we model

$$\gamma_m \sim \text{Gamma}(a_{\gamma_m},\, b_{\gamma_m}),$$

and

$$\gamma_{m+d}\,|\gamma_m \sim \text{Gamma}\left(\nu,\, \frac{\nu}{S(d)\,\gamma_m}\right),$$

for $d \in \mathbb{Z} \setminus \{0\}$. Then under these assumptions it follows that

$$\mathbb{E}\left[\gamma_{m+d}\,|\gamma_m\right] = S(d)\,\gamma_m,$$

and

$$\text{Var}\left[\gamma_{m+d}\,|\gamma_m\right] = \frac{1}{\nu}[S(d)\gamma_m]^2.$$

Thus S controls the amount of decay in the mean that the fertile window day coefficients undergo in comparison to the peak fertility day, while ν controls the variance from the decay that the model permits. Furthermore, in practice we will not know exactly which day will be the peak intensity day in the fertile window, so we allow m to be a random variable and define μ to be the value of the peak intensity day. More precisely, we define

$$\mu \,|\, m = \gamma_m,$$

where m is the index of the coefficient corresponding the peak intensity day. In Section 4.3.1.1 we discuss the selection of S, and in Sections 4.3.1.2, 4.3.1.3, and 4.3.1.4 we discuss the selection of μ, ν, and m.

4.3.1.1 Specification of the fertile window decay function S

We define the decay function S to be any function satisfying $S(-\infty) = S(\infty) = 0$ and $S(0) = 1$. In this paper we propose using a parametric function based on the skew-normal distribution, with values for the parameters based on the existing estimates of the probabilities of pregnancy relative to the timing of intercourse during the fertile window. We choose to use a parametric distribution so that researchers can easily extend the model by equipping the distributions parameters with prior distributions, however, in this paper we find it sufficient to consider the values of these parameters to be fixed quantities. Our reasoning for this is that because we have a limited number of days in the fertile window from which to infer the distributional parameters, it is questionable whether the added flexibility in the model is worth additional model complexity. Furthermore, by making ν a random variable, we can allow the model to deviate from the decay curve as much as the data dictates by decreasing the value of ν.

We view the skew-normal distribution as being a natural choice to base the decay function from among the class of parametric distributions, due to its unimodal form and its ability to take on a non-symmetric shape. We note that although selecting a decay curve with a unimodal form encourages a unimodal form in the regression coefficients, it does not strictly enforce it. We believe that this is ideal in the sense that we can inform the model based on our past experience and biological understanding, yet still allow the model freedom to deviate from this form if the data dictates it.

The skew-normal distribution has three parameters, a location parameter, a variance parameter, and a skew parameter. In order to select these parameters, we perform the following elicitation scheme. First, we consider the existing estimates of the probabilities of pregnancy relative to the timing of intercourse during the fertile window, and center and scale these estimates so that the peak intercourse days coincide and the scaled probability for the peak day takes a value of 1. Then we find the values for the skew-normal parameters that minimize the sum of the squared-error loss for the transformed historical data.

In Figure 4.1 we display the elicited decay curve. The values obtained from the elicitation scheme are a shift parameter of approximately -1.32, a variance of 2.68, and a skew parameter of -2.65 (and then the distribution is scaled so that the mode takes a value of 1).

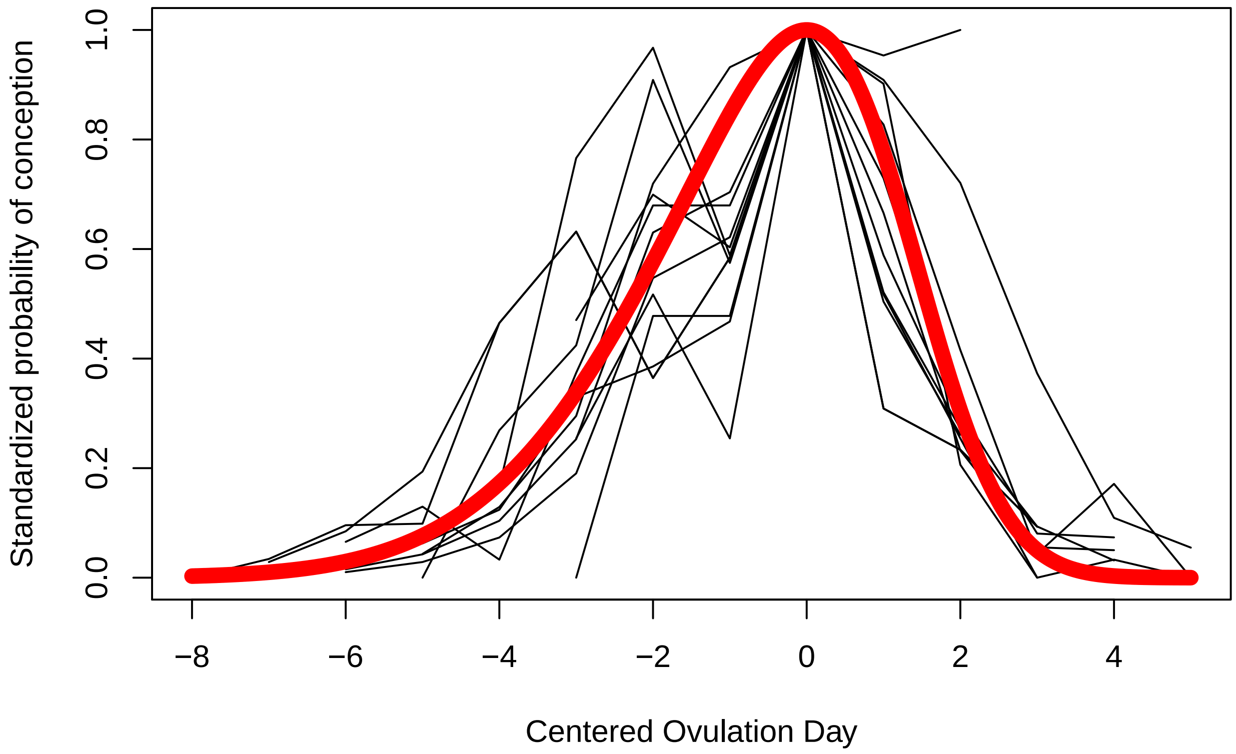


Figure 4.1: The elicited decay curve

4.3.1.2 Specification of the peak intensity parameter μ

We equip the peak intensity μ with a gamma prior distribution, and use the empirical mean obtained from the existing literature meta-analysis as the target mean for the prior distribution for the peak intensity parameter μ, in the sense that we select hyperparameters a_μ and b_μ that result in the desired mean. For the meta-analysis we obtained a target mean coefficient value of 0.34, and for the hyperparameters we chose values of $a_\mu = 0.34/0.2$ and $b_\mu = 1/0.2$, resulting in a unimodal distribution with a standard deviation of approximately 0.26. These hyperparameters were selected to find values that would achieve a flexible standard deviation as well as maintain a unimodal form.

4.3.1.3 Specification of the coefficient variance parameter ν

For the coefficient variance parameters we selected a gamma prior distribution with a hyperparameters $a_\nu = 25/8$ and $b_\nu = 1/8$ which results in a unimodal distribution with a mean of 25 and a standard deviation of approximately 14.14. So, for example when ν has a value of 25, $S(d)$ has a value of 1, and μ has a value of 0.34, then the standard deviation of γ_{m+d} is 0.68 (for a mean value of 0.34). The standard deviation is smaller at the tails of the decay curve, so for example, when ν has a value of 25, $S(d)$ has a value of 0.2, and μ has a value of 0.34, then the standard deviation of γ_{m+d} is 0.0136 (for a mean value of 0.068).

4.3.1.4 Specification of the peak intensity day parameter m

We propose modeling the peak intensity day using a multinomial distribution, with fixed prior probabilities chosen by the researcher based on their level of confidence in the scheme used to identify the ovulation day for the data in hand. For example, when ovulation day is determined using a relatively accurate measure such as an ovulation predictor kit in a preponderance of the data, then a researcher may want to place the bulk of the mass in the three or four days around the ovulation day. One the other hand, when the ovulation day is primarily identified by less accurate methods such as by counting backward from the next occurrence of menstrual bleeding, then a researcher may wish to place relatively flat prior probabilities on the distribution.

For the simulations and real data analysis, we used multinomial prior probabilities of 0.20, 0.25, and 0.20 for the three days that we most strongly expect to be the peak intensity day, and distributed the remaining mass among the other remaining days.

4.3.2 Updating step for the fertile window day coefficients and prior parameters

The full conditional distributions for the fertile window day coefficients and for the prior parameters do not have a closed-form solution (with the exception of the peak intensity day parameter m), so we use a Metropolis step to update each of the parameters. The full derivation of the updating steps is provided in the supplementary materials.

4.3.2.1 Proposal distributions

For each of the updating steps, we use the absolute value of a uniform distribution as the proposal distribution, with tuning parameters chosen to yield between 20% and 40% acceptance rates.

4.4 Missing data imputation

One of the potential limitations of the day-specific probability models proposed in Dunson and Stanford (2005) is that it doesn't have a mechanism for incorporating missing data. As a result, any cycles that have a missing value have to be removed from the data, which can obviously result in a loss of efficiency as the amount of missingness increases.

In our experience, the intercourse status for a given day is the most commonly missing variable. Consider the worst-case scenario where missingness for this variable is independent – then 10% missingness results in only 48% of cycles being retained for a 7-day fertile window on average, while a 20% missingness results in only 21% of the cycles being retained for a 7-day fertile window. Fortunately, the missingness of intercourse status tends to be correlated within a cycle so that cycles often have either no missing or multiple missing values, but regardless data is expensive to collect and we want to use as much of it as possible.

Missing baseline variables such as e.g. gravidity status have a different issue that a missing value causes all of the participant's cycles to be thrown out. Although we would expect it to be less common to see missingness in this type of variable, once again this kind of attrition can add up which we obviously want to avoid.

4.4.1 Gibbs step for missing X_{ijk}

We propose an autoregressive prior for the distribution of a missing intercourse indicator X_{ijk}. This is motivated by our empirical evidence that the best predictor of intercourse for the i, j, k-th day is whether intercourse occurred on the previous day. We assume that

$$\begin{aligned}
&\mathbb{P}\left(X_{ijk} = 1 \,\middle|\, \boldsymbol{X}_{-ijk}, \boldsymbol{U}\right) \\
&\qquad = \mathbb{P}\left(X_{ijk} = 1 \,\middle|\, X_{ij,k-1}, \boldsymbol{U}\right) \\
&\qquad = \frac{\exp\left\{(X_{ij,k-1}, \boldsymbol{U}_{ijk})^T \boldsymbol{\tau}\right\}}{1 + \exp\left\{(X_{ij,k-1}, \boldsymbol{U}_{ijk})^T \boldsymbol{\tau}\right\}},
\end{aligned}$$

for $\boldsymbol{\tau}$ a known vector of regression coefficients. In other words, the prior probability of intercourse on the i, j, k-th day is a logistic regression model. In practice, we don't know the true values of $\boldsymbol{\tau}$ *a priori* so we use an empirical Bayes approach and estimate it from the nonmissing entries in the data.

Under this assumption, we obtain the following full conditional distribution for X_{ijk} when $k < K$ (i.e. not the last day in the fertile window). When $k = K$ the full conditional distribution of X_{ijk} is similar to that of (4.1), except without the $X_{ij,k+1}$ terms.

$$\mathbb{P}\Big(X_{ijk} = 1 \,\Big|\, \boldsymbol{Y}, \boldsymbol{W}, \boldsymbol{\gamma}, \boldsymbol{\xi}, \phi, \boldsymbol{X}_{-ijk}, \boldsymbol{U}, \text{data}\Big) \tag{4.1}$$

$$= \frac{\pi\Big(W_{ijk} \mid \boldsymbol{\gamma}, \xi_i, X_{ijk} = 1, \boldsymbol{X}_{-ijk}, \boldsymbol{U}, \text{data}\Big)\, \mathbb{P}\left(X_{ij,k+1} \,|\, X_{ijk} = 1, \boldsymbol{U}\right) \mathbb{P}\left(X_{ijk} = 1 \,|\, X_{ij,k-1}, \boldsymbol{U}\right)}{\sum_{t=0}^{1} \pi\Big(W_{ijk} \mid \boldsymbol{\gamma}, \xi_i, X_{ijk} = t, \boldsymbol{X}_{-ijk}, \boldsymbol{U}, \text{data}\Big)\, \mathbb{P}\left(X_{ij,k+1} \,|\, X_{ijk} = t, \boldsymbol{U}\right) \mathbb{P}\left(X_{ijk} = t \,|\, X_{ij,k-1}, \boldsymbol{U}\right)}$$

We then fully specify the joint distribution as follows. We assume that

$$\begin{aligned}\mathbb{P}\left(\boldsymbol{X} \,|\, \boldsymbol{U}\right) &= \prod_{i,j} \mathbb{P}\left(X_{ij1}, \ldots, X_{ijK} \mid \boldsymbol{U}\right) \\ &= \prod_{i,j} \Big\{\mathbb{P}\left(X_{ijK} \,|\, X_{ij,K-1}, \boldsymbol{U}\right) \cdots \mathbb{P}\left(X_{ij2} \,|\, X_{ij1}, \boldsymbol{U}\right) \mathbb{P}\left(X_{ij1} \,|\, \boldsymbol{U}\right)\Big\}.\end{aligned}$$

In Section 4.6 we investigate the intercourse status missingness in the Time to Conceive dataset, and fit the autoregressive empirical Bayes model for the nonmissing data.

4.4.2 Gibbs step for missing covariates

Missingness in the covariates can come in the form of a day-specific variable, a cycle-specific variable, or a baseline variable. In the interest of brevity we characterize the updating step for a missing continuous covariate here, and in the supplementary materials we describe the updating step for missing cycle-specific or baseline covariates.

We assume that the prior distributions for the covariates are mutually independent. Although we acknowledge that this is not a realistic assumption for every setting, our contention is that without study-specific knowledge of the variables involved, it is difficult to specify a reasonable joint distribution model. We expect this to be a reasonable approximation for what is ideally a relatively small amount of missingness in the data.

The model for missing continuous covariates and missing categorical covariates differ slightly, as we will cover in the following sections.

4.4.2.1 Gibbs step for missing continuous covariate

Let U_{ijkh} be a missing day-specific covariate, then we use a Metropolis step to update U_{ijkh}. Under the previously stated assumptions, we have

$$\begin{aligned}
&\pi\Big(U_{ijkh} \,\Big|\, \boldsymbol{Y}, \boldsymbol{W}, \boldsymbol{\gamma}, \boldsymbol{\xi}, \phi, \boldsymbol{U}_{-ijkh}\Big) \\
&\qquad = \frac{\pi\Big(\boldsymbol{Y}, \boldsymbol{W}, \boldsymbol{\gamma}, \boldsymbol{\xi}, \phi, U_{ijkh}, \boldsymbol{U}_{-ijkh}\Big)}{\int \pi\Big(\boldsymbol{Y}, \boldsymbol{W}, \boldsymbol{\gamma}, \boldsymbol{\xi}, \phi, U_{ijkh}, \boldsymbol{U}_{-ijkh}\Big)\, dU_{ijkh}} \\
&\qquad \propto \frac{\pi\Big(\boldsymbol{W}_i \,\Big|\, \boldsymbol{\gamma}, \xi_i, \boldsymbol{X}, U_{ijkh}, \boldsymbol{U}_{-ijkh}\Big)\, \mathbb{P}\left(X_{ijk} \,|\, U_{ijkh}\right)\, \pi(U_{ijkh})}{\int \pi\Big(\boldsymbol{W}_i \,\Big|\, \boldsymbol{\gamma}, \xi_i, \boldsymbol{X}, U_{ijkh}, \boldsymbol{U}_{-ijkh}\Big)\, \mathbb{P}\left(X_{ijk} \,|\, U_{ijkh}\right)\, \pi(U_{ijkh})\, dU_{ijkh}},
\end{aligned}$$

where the $\mathbb{P}\left(\boldsymbol{X}_{ijk} \,|\, U_{ijk}\right)$ term is a constant if X_{ijk} is nonmissing for the i^{th} subject. It follows that the acceptance ratio r is given by

$$\begin{aligned}
&r_{\text{ day-specific}} \\
&\qquad = \frac{\pi\Big(U_{ijkh} = u^* \,\Big|\, \boldsymbol{Y}, \boldsymbol{W}, \boldsymbol{\gamma}, \boldsymbol{\xi}, \phi, \boldsymbol{U}_{-ijkh}\Big)}{\pi\Big(U_{ijkh} = u^{(s)} \,\Big|\, \boldsymbol{Y}, \boldsymbol{W}, \boldsymbol{\gamma}, \boldsymbol{\xi}, \phi, \boldsymbol{U}_{-ijkh}\Big)} \\
&\qquad = \frac{\pi\Big(\boldsymbol{W}_i \,\Big|\, \boldsymbol{\gamma}, \xi_i, \boldsymbol{X}, U_{ijkh} = u^*, \boldsymbol{U}_{-ijkh}\Big)\, \mathbb{P}\left(X_{ijk} \,|\, U_{ijkh} = u^*\right)\, \pi(U_{ijkh} = u^*)}{\pi\Big(\boldsymbol{W}_i \,\Big|\, \boldsymbol{\gamma}, \xi_i, \boldsymbol{X}, U_{ijkh} = u^{(s)}, \boldsymbol{U}_{-ijkh}\Big)\, \mathbb{P}\left(X_{ijk} \,|\, U_{ijkh} = u^{(s)}\right)\, \pi(U_{ijkh} = u^{(s)}).}
\end{aligned}$$

Additionally, we also need to specify a prior distribution for the covariates. For this paper we used a conjugate prior Normal distribution for the imputed covariate. In more detail, we specify that

$$\begin{aligned}
U_{ijkh} \,|\, \zeta, \kappa &\overset{i.i.d.}{\sim} N(\zeta, \kappa), \qquad \text{for all } i, j, k, h \\
\zeta \,|\, \kappa, \{U_{ijkh}\} &\sim N(\zeta_0, \kappa), \\
\kappa \,|\, \{U_{ijkh}\} &\sim \text{Gamma}(a_\kappa, b_\kappa),
\end{aligned}$$

where $\{U_{ijkh}\}$ is taken to mean the set of all missing h^{th} covariates. We take an Empirical Bayes approach when specifying the hyperparameters. We set ζ_0 to be the empirical mean of the nonmissing h^{th} covariates, and select a_κ, and b_κ so that the prior distribution's mean is equal to the sample variance of the nonmissing

h^{th} covariates and let the distribution have a noninformative variance. We use a uniform distribution as the proposal distribution for the missing covariates.

4.4.2.2 Gibbs step for missing categorical covariate

Let U_{ijk} be a day-specific categorical variable with corresponding design matrix dummy variables $(U_{ijk\ell_1}, \dots U_{ijk\ell_r})$. Then

$$\begin{aligned}
&\mathbb{P}\Big(U_{ijk} = u \,\big|\, \boldsymbol{Y}, \boldsymbol{W}, \boldsymbol{\gamma}, \boldsymbol{\xi}, \phi, \boldsymbol{U}_{-ijk}\Big) \\
&\quad = \frac{\pi\Big(\boldsymbol{Y}, \boldsymbol{W}, \boldsymbol{\gamma}, \boldsymbol{\xi}, \phi, U_{ijk} = u, \boldsymbol{U}_{-ijk}\Big)}{\sum_t \pi\Big(\boldsymbol{Y}, \boldsymbol{W}, \boldsymbol{\gamma}, \boldsymbol{\xi}, \phi, U_{ijk} = t, \boldsymbol{U}_{-ijk}\Big)} \\
&\quad = \frac{\mathbb{P}\Big(W_{ijk} \,\big|\, \boldsymbol{\gamma}, \xi_i, \boldsymbol{X}, U_{ijk} = u, \boldsymbol{U}_{-ijk}\Big)\,\mathbb{P}\left(X_{ijk} \,|\, X_{ij,k-1}, U_{ijk} = u\right)\mathbb{P}\left(U_{ijk} = u\right)}{\sum_t \mathbb{P}\Big(W_{ijk} \,\big|\, \boldsymbol{\gamma}, \xi_i, \boldsymbol{X}, U_{ijk} = t, \boldsymbol{U}_{-ijk}\Big)\,\mathbb{P}\left(X_{ijk} \,|\, X_{ij,k-1}, U_{ijk} = t\right)\mathbb{P}\left(U_{ijk} = t\right)}.
\end{aligned}$$

We specify the prior distribution to be conjugate multinomial-Dirichlet distributed. That is to say we let

$$\begin{aligned}
U_{ijkh} \,|\, \boldsymbol{\pi}_h &\sim \text{multinomial}(\boldsymbol{\pi}_h) \\
\boldsymbol{\pi}_h &\sim \text{Dirichlet}(\boldsymbol{a}_h)
\end{aligned}$$

We use an empirical Bayes approach to specify the mean of the prior distributions of the $\boldsymbol{a}_h$.

4.5 Simulation studies

4.5.1 Comparing the peak-intensity model

In order to compare the peak-intensity decay model, we compare the peak-intensity decay model with two versions of the day-specific probabilities model proposed in Dunson and Stanford (2005). In the first version of the day-specific probabilities model, we use the data that we generated in 4.3.1.1 to provide informative priors to the model coefficients. Then, for the second specification of the model we used uninformative priors.

In order to compare these various models, we considered a number of different settings. In terms of the simulated data settings, we consider fertile window sizes of both five days and seven days. In terms of the sample size we considered sizes of 100, 200, and 300 participants. For the coefficients corresponding to the fertile window days, we consider three settings.

In the first setting, we set the coefficients to be the same as the decay that we derived in Section 4.3.1.1, so that both the peak-intensity decay model and the informative priors day-specific probabilites models are correctly specified. In another setting, we set the peak-intensity of the coefficients corresponding to the fertile window to be scaled from that of the prior distributions, and we shifted the location of the peak-intensity day. We would expect that the peak-intensity decay model would be able to recover these parameters more readily than the incorrectly specified informative priors day-specific probabilities model. Finally, we also included a model in which the decay of the coefficients corresponding to the fertile window days is different than what was derived in 4.3.1.1, so that both the peak-intensity decay model and the informative priors day-specific probabilities model are incorrectly specified.

We present the results from a pair of simulation studies in what follows, and in the Supplementary Materials we present an extended set of simulation results. For each of the simulation studies we are describing the average results over 10,000 simulations.

4.5.1.1 Correctly specified mean prior distribution setting

In Table B.2 we compare the posterior distributions for the fertile window day coefficients for a simulated set of data such that the coefficients used for the data-generating mechanism are equal to the means of the prior distributions for the peak-intensity model and the informative prior version of the day-specific probabilities model. Each dataset was generated with 5 fertile window days and 200 subjects.

For this simulation setting, we see that for each of the coefficients corresponding to the fertile window days that the average squared error between the posterior mean and the true coefficient value is smallest for the peak intensity model compared to the two prior specifications of the day-specific probabilities model. Furthermore, the standard deviation of the posterior mean distributions across all of the simulations is smaller for each of the coefficients for the peak intensity model, and the average width of the 95% credible region is smaller than the other model parameterizations. We attribute these reductions in variance in the posterior distributions for the peak intensity model to the induced correlation structure between the coefficients in the prior distribution. This allows for information sharing between the coefficients that results in a gain of efficiency compared to the other models.

On the other hand, we see that the coefficient's posterior means for the peak intensity model tend to have a slightly larger average difference magnitude than the other models. However, the largest magnitude of average difference between the posterior mean and the true coefficient value is 0.016 (for fertile window day 3), which we find to be an acceptably small difference. We attribute this behavior to the form of the peak-intensity model. For this model, the posterior mean is the mean of the mixture distribution comprised from the posterior distributions conditional on the peak intensity locations. When the peak intensity day location is shifted towards a given fertile window day, then the coefficient corresponding to this day will be stochastically larger, while the converse is true when the peak intensity day location is shifted away from that same fertile window day. When we look at the posterior distribution for the peak intensity day location, we see that the day 5 of the fertile window has the most mass outside of the true peak intensity day (i.e. day 4). As a result, the posterior means for the coefficients corresponding to the fertile window days before the peak intensity day have a diminished unconditional mean, while the coefficient corresponding to the day after the peak intensity day is increased.

It is interesting that day 5 has the largest posterior probability of being the peak intensity day outside of the true peak intensity day. In our experience, the model tends to favor the "shorter side" of the fertile window, meaning the side of the fertile window that has less days before or after the true peak intensity day. The reason for this behavior is a byproduct of the induced correlation between the fertile window day coefficients. In the posterior distribution conditional on day 5 being the peak intensity day, when the coefficient values are larger than the conditional posterior mean, then they can fit the true coefficients quite well, with only the exception of day 5 having an inaccurate value. On the other hand, for the posterior distribution conditional on day 3 being the peak intensity day, then when the correlation structure of the coefficients holds, then about half of the fertile window days have an inaccurate value, regardless of the direction of the correlation relative to the mean. As expected, we can see that the posterior mean for the peak intensity magnitude is slightly larger than the true value.

When the size of the fertile window is larger, such as seven days compared to five, then a shift of one day for the peak intensity day causes less of an imbalance, and as we would expect the model does not favor the shorter side as strongly. Furthermore, as the sample size increases, then the posterior distribution increasing places the bulk of the mass on the true peak intensity day.

4.5.1.2 Shifted and scaled mean prior distribution setting

In Table B.15, we compare the posterior distributions for the fertile window day coefficients for a simulated set of data such that the coefficients used for the data-generating mechanism are a shifted and scaled version of the means of the prior distributions for the peak-intensity model and the informative prior version of the day-specific probabilities model. More precisely, the mean of the prior distribution of the d^{th} day relative to the peak-intensity day is given by $\mu S(d)$ (where μ and S are chosen as described in Section 4.3.1), while the coefficient used for the d^{th} day for the data-generating mechanism is given by $\mu' S(d + a)$. In particular, the coefficients were scaled by a factor of 1.2, and the peak-intensity day was shifted two days to the left. Each dataset was generated with 7 fertile window days and 300 subjects.

This data-generating mechanism is particularly problematic for the day-specific probabilities model with informative priors, those priors are misspecified. This prior specification expects the coefficient corresponding to day 5 to have the greatest magnitude, while in fact it is the coefficient corresponding to day 3 that has the greatest magnitude. Because of this, the posterior means for the coefficients to the left of day 5 are smaller than the values used for the data simulations, while the posterior means for the coefficients to the right of day 5 are larger than the values used for the data simulations. The non-informative prior specification of the day-specific probabilities model on the other hand has posterior means that are closer to true coefficient values, however it still suffers from relatively large variances for the posterior distributions.

The peak-intensity model fares relatively well in this setting, as we would expect. The model has parameters for the magnitude of the scaling as well as for the peak intensity day location that can each be updated to reflect the data. We can see that the variance for this model is significantly lower than for the other models. For example, the width of the 95% credible region for the peak day coefficient is 0.221 for the peak-intensity model, while it is 0.274 and 0.275 for the informative and non-informative day-specific probabilities model, respectively.

We note that the posterior mean for the peak intensity magnitude is smaller than the peak intensity used for the data-generating mechanism. We find that although the peak intensity parameter is able to adapt to the data, it can be relatively slow to do so since its posterior distribution is based on the fertile window day coefficients which only has a small number of data points, and is only indirectly affected by the sample size through these coefficients. So since the mean of the prior distribution for the peak intensity magnitude is smaller than that what was used for the data simulation, (0.340 compared to 0.408), we see that a sample size of 300 is not yet sufficient for the posterior distribution to be fully centered around the true value.

4.5.2 Imputing missing intercourse status

4.5.2.1 Missing intercourse status fertile window model only

In Table 4.3 we compare the posterior distributions for the fertile window day coefficients for a simulated set of data with missing values inserted into the missing sexual intercourse status values. Each dataset was generated with 7 fertile window days, and to insert the missing values, we randomly changed at least one each sexual intercourse status value in a given cycle to missing with probability 0.52, independent of all other cycles (the value of 0.52 was chosen to be comparable with the probability of missing at least value if each day's intercourse was missing with probability 0.1 independent of any other days). In our experience missing sexual intercourse status values in a cycle are highly correlated. If a cycle was chosen to have missing sexual intercourse status values for a given simulation, then we inserted between 1 and 7 days of missing values with a uniform probability.

Each dataset was generated with 300 subjects. We simulated the data such that the coefficients used for the data-generating mechanism are equal to the means of the prior distributions for the peak-intensity model and the informative prior version of the day-specific probabilities model. We can see that on average 297 subjects had at least 1 cycle with sexual intercourse (or a missing value for sexual intercourse status) during the fertile window, for a total of 564 menstrual cycles on average. On the other hand, when we didn't impute missing sexual intercourse status, we were only able to retain an average of 186 subjects and 261 menstrual cycles.

When comparing the fixed informative models, we see that the posterior means had similar measures for mean squared error and standard deviation. However the variance of the posterior distribution was greatly reduced for the model with imputation. For example, the posterior distribution for day 5 (the peak intensity day) for the model with imputation had an average width of 0.303 for the 95% credible region, while the width of the credible region for the same coefficient in the non-imputation model had a width of 0.350. For the day 4 coefficient with true value 0.288, the model with imputation had an average width of 0.288 for the 95% credible region, while the width of the credible region for the same coefficient in the non-imputation model had a width of 0.329.

Table 4.1: Correctly specified prior parameters, 5 fertile window days, $n = 200$

day	true	mean	bias	MSE	Std.	CR
				Peak-intensity		
Day -3	0.115	0.105	-0.010	0.001	0.031	0.121
Day -2	0.196	0.182	-0.014	0.001	0.042	0.162
Day -1	0.288	0.272	-0.016	0.002	0.050	0.195
Day 0	0.340	0.331	-0.009	0.002	0.056	0.218
Day 1	0.264	0.273	0.009	0.004	0.064	0.251
				Fixed informative		
Day -3	0.115	0.102	-0.013	0.002	0.048	0.177
Day -2	0.196	0.185	-0.011	0.003	0.057	0.222
Day -1	0.288	0.279	-0.009	0.004	0.066	0.258
Day 0	0.340	0.336	-0.004	0.004	0.071	0.278
Day 1	0.264	0.266	0.002	0.004	0.066	0.254
				Non-informative		
Day -3	0.115	0.120	0.005	0.003	0.053	0.198
Day -2	0.196	0.188	-0.008	0.004	0.061	0.234
Day -1	0.288	0.273	-0.015	0.005	0.069	0.269
Day 0	0.340	0.326	-0.014	0.006	0.074	0.289
Day 1	0.264	0.263	-0.001	0.004	0.069	0.266

Posterior distribution for the peak intensity

true	mean	bias	50% CR	90% CR
0.340	0.354	0.014	0.075	0.145

Posterior distribution for the peak
intensity location

Day -3	Day -2	Day -1	Day 0	Day 1
0.000	0.000	0.080	0.633	0.287

Table 4.2: Scaled and shifted prior parameters, 7 fertile window days, $n = 300$

day	true	mean	bias	MSE	Std.	CR
			Peak-intensity			
Day -2	0.235	0.232	-0.003	0.001	0.042	0.166
Day -1	0.346	0.341	-0.005	0.002	0.052	0.205
Day 0	0.408	0.402	-0.006	0.002	0.056	0.221
Day 1	0.317	0.304	-0.013	0.002	0.049	0.192
Day 2	0.124	0.118	-0.006	0.000	0.027	0.107
Day 3	0.020	0.019	-0.001	0.000	0.008	0.023
Day 4	0.002	0.002	0.000	0.000	0.003	0.002
			Fixed informative			
Day -2	0.235	0.190	-0.045	0.006	0.058	0.222
Day -1	0.346	0.309	-0.037	0.005	0.066	0.257
Day 0	0.408	0.378	-0.030	0.005	0.070	0.274
Day 1	0.317	0.304	-0.013	0.003	0.064	0.251
Day 2	0.124	0.160	0.036	0.003	0.051	0.196
Day 3	0.020	0.067	0.047	0.003	0.037	0.136
Day 4	0.002	0.004	0.002	0.000	0.013	0.029
			Non-informative			
Day -2	0.235	0.229	-0.006	0.003	0.062	0.238
Day -1	0.346	0.335	-0.011	0.004	0.069	0.267
Day 0	0.408	0.395	-0.013	0.005	0.073	0.284
Day 1	0.317	0.300	-0.017	0.004	0.066	0.259
Day 2	0.124	0.126	0.002	0.002	0.051	0.195
Day 3	0.020	0.031	0.011	0.001	0.030	0.103
Day 4	0.002	0.016	0.014	0.000	0.024	0.076

Posterior distribution for the peak intensity

true	mean	bias	50% CR	90% CR
0.408	0.397	-0.011	0.064	0.124

Posterior distribution for the peak
intensity location

Day -2	Day -1	Day 0	Day 1	Day 2	Day 3	Day 4
0.000	0.032	0.962	0.006	0.000	0.000	0.000

When comparing the peak-intensity model with imputation to the fixed informative model with imputation, we again see that the posterior mean for the peak-intensity model is slightly further on average than for the fixed intensity model. However, the mean squared error and standard deviation of the posterior mean smaller for the peak-intensity model (for example, a MSE of 0.002 for the peak-intensity model for day 5 compared to a MSE of 0.004 for the fixed informative model). Furthermore, the width of the coefficient's 95% credible regions is far smaller on average than for the fixed informative model. For example, for the peak intensity day, the average width of the credible region for the peak-insensity model was 0.218 compared to a width of 0.303 for the fixed informative model (and an average width of 0.350 for the non-imputation model).

4.5.2.2 Missing intercourse status fertile window plus covariates model

In Table 4.4 we again compared the posterior distributions for the fertile window day coefficients for a simulated set of data with missing values inserted into the missing sexual intercourse status values, however this time we also included a set of covariates into the data-generating mechanism. Each dataset was generated with 5 fertile window days, and to insert the missing values, we randomly changed each sexual intercourse status value to missing with probability 0.1.

Each dataset was generated with 300 subjects. We simulated the data such that the coefficients used for the data-generating mechanism are equal to the means of the prior distributions for the peak-intensity model and the informative prior version of the day-specific probabilities model. We included six categorical covariates which we have labeled to emulate dummy-encoded coefficients corresponding to covariates for age, BMI and gravidity status. The hyperparameters for gamma distribution for the non-fertile window day coefficients were chosen to have a value of 2 for both the shape and the rate parameters so as to provide a unimodal distribution with na posterior mean of 1 (i.e. no effect on the model) such that nearly all of the mass is between 0 and 3.

We can see that on average 293 subjects had at least 1 cycle with sexual intercourse (or a missing value for sexual intercourse status) during the fertile window, for a total of 537 cycles on average. On the other hand, when we didn't impute missing sexual intercourse status, we were only able to retain an average of 204 subjects and 293 menstrual cycles.

Both the peak-intensity model and fixed informative model with missing data imputation tended to have smaller mean squared errors and credible region widths than the fixed informative model without missing data imputation. For example, when comparing the coefficient corresponding to Day 0, the peak-intensity model credible region width is 0.248 on average compared to 0.342 for the non-imputation model, and when comparing coefficient corresponding to Age 36-38 status, the mean squared error for the fixed informative model with imputation is 0.036 compared to 0.068 for the model without imputation.

We can see that this scenario is similar to the previous missing-data scenario in the sense that the coefficients corresponding to the fertile window days for the peak-intensity model tend to have a posterior mean that is smaller than the true value used to generate the data, however the effect is more pronounced in this case. The average mean squared errors and posterior variances for the coefficients corresponding to the fertile window days still tends to be smaller for the peak-intensity model than for the fixed-informative models for either with or without missing data imputation. On the other hand, the posterior mean for the coefficients corresponding to the non-fertile window days tends to be larger for the peak-intensity model with missing data imputation than for the fixed informative model with missing data imputation. In general, we expect the mass of the posterior distributions to tend towards 1 since we specified a prior distribution with a mean of 1. However, we view the addition of the covariates as adding more complexity to the model, and therefore causing decreased accuracy in our posterior estimate of the peak intensity day. This in turn causes decreased posterior means for the coefficients corresponding to the fertile window day, and the posterior distributions for the non-fertile window day coefficients are stochastically larger as a result.

Because of this effect, there is something of a trade-off when the amount of posterior uncertainly increases due to the amount of covariates in the model or the quantity of missingness. In short, when the posterior variance of the fertile window day coefficients is the primary concern then we recommend using the peak-intensity model, whereas if the posterior distribution of the covariates is the primary scientific interest then we recommend using the fixed informative model.

4.6 Real Data Analysis

To asses the performance of the methodology proposed in this paper, we consider the Time to Conceive (TTC) cohort, a prospective time-to-pregnancy study. This study included English-speaking women between 29 and 44 years of ages who had been trying to conceive for 3 months or less. The study did not administer biological measurements to ascertain the ovulation day, so in order to use the most accurate measures of ovulation possible, wer restricted our attention to cycles in which ovulation was determined through a positive

Table 4.3: Imputing sexual intercourse status without covariates, 7 fertile window days, correlated missingness mechanism, $n = 300$

day	true	mean	bias	MSE	Std.	CR
			Peak-intensity			
Day -4	0.059	0.056	-0.003	0.000	0.019	0.074
Day -3	0.115	0.110	-0.005	0.001	0.031	0.121
Day -2	0.196	0.188	-0.008	0.001	0.043	0.167
Day -1	0.288	0.278	-0.010	0.002	0.052	0.204
Day 0	0.340	0.325	-0.015	0.002	0.055	0.218
Day 1	0.264	0.255	-0.009	0.003	0.055	0.215
Day 2	0.104	0.111	0.007	0.002	0.044	0.169
			Fixed informative			
Day -4	0.059	0.031	-0.028	0.003	0.024	0.075
Day -3	0.115	0.108	-0.007	0.003	0.055	0.203
Day -2	0.196	0.197	0.001	0.003	0.065	0.253
Day -1	0.288	0.295	0.007	0.004	0.074	0.288
Day 0	0.340	0.342	0.002	0.004	0.077	0.303
Day 1	0.264	0.271	0.007	0.004	0.071	0.277
Day 2	0.104	0.092	-0.012	0.003	0.050	0.185
			Non-informative			
Day -4	0.059	0.029	-0.030	0.003	0.025	0.078
Day -3	0.115	0.101	-0.014	0.004	0.057	0.213
Day -2	0.196	0.191	-0.005	0.004	0.073	0.281
Day -1	0.288	0.288	0.000	0.005	0.085	0.329
Day 0	0.340	0.337	-0.003	0.006	0.090	0.350
Day 1	0.264	0.265	0.001	0.005	0.081	0.315
Day 2	0.104	0.087	-0.017	0.004	0.052	0.191

Average number of subjects and cycles (imputation)

Subjects	Cycles
297	564

Average number of subjects and cycles (non-imputation)

Subjects	Cycles
186	261

Table 4.4: Imputing sexual intercourse status with covariates, 5 fertile window days, independent missingness mechanism, $n = 300$

day	true	mean	bias	MSE	Std.	CR
			Peak-intensity			
Day -3	0.115	0.099	-0.016	0.001	0.029	0.111
Day -2	0.196	0.174	-0.022	0.002	0.043	0.166
Day -1	0.288	0.257	-0.031	0.003	0.055	0.216
Day 0	0.340	0.308	-0.032	0.004	0.064	0.248
Day 1	0.264	0.249	-0.015	0.003	0.060	0.235
Age 30-35	0.923	1.040	0.117	0.049	0.170	0.665
Age 36-38	0.651	0.741	0.090	0.049	0.191	0.742
Age 38+	0.357	0.427	0.070	0.031	0.126	0.492
BMI Overweight	0.803	0.864	0.062	0.030	0.162	0.635
BMI Obese	0.625	0.683	0.058	0.022	0.147	0.571
Gravid Yes	2.226	2.377	0.152	0.167	0.368	1.437
			Fixed informative			
Day -3	0.115	0.117	0.002	0.002	0.044	0.170
Day -2	0.196	0.204	0.008	0.002	0.059	0.228
Day -1	0.288	0.288	0.000	0.004	0.069	0.268
Day 0	0.340	0.341	0.001	0.004	0.076	0.295
Day 1	0.264	0.268	0.004	0.003	0.064	0.251
Age 30-35	0.923	0.977	0.054	0.031	0.154	0.602
Age 36-38	0.651	0.697	0.047	0.036	0.177	0.690
Age 38+	0.357	0.402	0.045	0.026	0.118	0.459
BMI Overweight	0.803	0.830	0.027	0.024	0.154	0.603
BMI Obese	0.625	0.656	0.031	0.019	0.140	0.545
Gravid Yes	2.226	2.214	-0.012	0.109	0.325	1.272
			Non-informative			
Day -3	0.115	0.117	0.002	0.002	0.051	0.196
Day -2	0.196	0.202	0.006	0.003	0.069	0.269
Day -1	0.288	0.285	-0.003	0.005	0.081	0.316
Day 0	0.340	0.326	-0.014	0.006	0.088	0.342
Day 1	0.264	0.266	0.002	0.004	0.077	0.300
Age 30-35	0.923	1.000	0.077	0.063	0.207	0.809
Age 36-38	0.651	0.733	0.083	0.068	0.238	0.923
Age 38+	0.357	0.422	0.065	0.046	0.157	0.608
BMI Overweight	0.803	0.853	0.050	0.047	0.205	0.798
BMI Obese	0.625	0.692	0.067	0.035	0.187	0.729
Gravid Yes	2.226	2.206	-0.020	0.140	0.410	1.603

result of an ovulation predictor kit. The fertile window was defined as extending four days before and four days after the ovulation day.

For the data analysis we considered the same three models as in Tables 4.3 and 4.4, namely the peak-intensity model, the fixed informative model, and the fixed non-informative model. In order to obtain a fair comparison only included cycles without any missing data in them. We also include categorical covariates for age, BMI, and gravidity status.

We show the results of analyzing the TTC dataset in Table 4.5. We can see that there are a total of 334 women with a total of 585 menstrual cycles included in the study. When considering the coefficients corresponding to the fertile window days, we see that for all three models the coefficients for the most part follow a unimodal form with the days most likely to result in conception occurring during the three days before ovulation. As discussed before, we see that the peak-intensity model tends to have lower posterior means than the fixed informative model due to the uncertainty in the location of the peak intensity day, with the benefit of having smaller posterior variances.

An interesting anomaly occurs for the fixed informative and non-informative models for the coefficient corresponding to fertile window day 4. For this day we have a sudden spike in the posterior mean, with a value of 0.100 for the fixed informative model, and a value of 0.250 for the fixed non-informative model. In comparison, the information borrowing across the days results in a posterior mean of 0.001 for the peak-intensity model for this coefficient. Since this unexpected spike is not a reasonable value for the data, we can speculate that this is some sort of misspecification of the data occuring during the data collection process. Whatever the cause, we can see that the peak-intensity model defends us against unrealistic posterior distributions such as these.

In each of the models, the posterior means for the age categories each have values below 1, which has the interpretation of these age categories having lower fecundability than for the reference group of ages 29-30. Next, for the overweight BMI status category, we observe a value of close to 1 for each of the models, which has the interpretation of no effect in the model in comparison to the reference group of underweight BMI and normal weight BMI combined (the underweight category is too small to be considered on its own). A BMI status of obese has a value below 1 for each of the models, indicating reduced levels of fecundability compared to the reference groups. And finally, having had a previous pregnancy indicates a higher ability to become pregnant than women who have never become pregnant.

Table 4.5: Analysis of the Time to Conceive data

	Peak-intensity			
Coef.	Mean	Std.	CR Low	CR High
			Peak-intensity	
Day -4	0.054	0.047	0.000	0.149
Day -3	0.145	0.050	0.057	0.253
Day -2	0.166	0.064	0.049	0.291
Day -1	0.238	0.059	0.144	0.357
Day 0	0.135	0.065	0.040	0.254
Day 1	0.037	0.048	0.000	0.112
Day 2	0.009	0.031	0.000	0.023
Day 3	0.002	0.019	0.000	0.002
Day 4	0.001	0.013	0.000	0.002
Age 31-35	0.932	0.225	0.590	1.464
Age 36-38	0.681	0.235	0.332	1.234
Age 39+	0.371	0.180	0.121	0.812
BMI Overweight	1.029	0.253	0.610	1.612
BMI Obese	0.832	0.248	0.435	1.389
Gravid Yes	1.454	0.291	0.969	2.097
			Fixed informative	
Day -4	0.060	0.058	0.000	0.205
Day -3	0.169	0.072	0.051	0.334
Day -2	0.194	0.072	0.079	0.362
Day -1	0.291	0.076	0.168	0.464
Day 0	0.154	0.061	0.056	0.294
Day 1	0.032	0.048	0.000	0.168
Day 2	0.003	0.019	0.000	0.033
Day 3	0.001	0.011	0.000	0.000
Day 4	0.100	0.146	0.000	0.480
Age 31-35	0.811	0.178	0.513	1.206
Age 36-38	0.612	0.209	0.295	1.097
Age 39+	0.322	0.156	0.108	0.701
BMI Overweight	0.990	0.250	0.579	1.560
BMI Obese	0.804	0.239	0.421	1.346
Gravid Yes	1.359	0.270	0.910	1.952
			Non-informative	
Day -4	0.099	0.069	0.008	0.265
Day -3	0.174	0.081	0.048	0.363
Day -2	0.183	0.083	0.053	0.374
Day -1	0.303	0.088	0.163	0.504
Day 0	0.145	0.072	0.037	0.314
Day 1	0.103	0.074	0.008	0.282
Day 2	0.091	0.071	0.006	0.271
Day 3	0.078	0.066	0.004	0.246
Day 4	0.250	0.136	0.043	0.575
Age 31-35	0.730	0.167	0.455	1.092
Age 36-38	0.561	0.197	0.267	1.019
Age 39+	0.276	0.136	0.092	0.612
BMI Overweight	0.956	0.247	0.552	1.534
BMI Obese	0.770	0.240	0.387	1.320
Gravid Yes	1.310	0.266	0.867	1.902

4.7 Conclusion

In this paper we propose a new model called the peak-intensity model for describing the contribution of sexual intercourse during a given fertile window day towards the likelihood of becoming pregnant during the menstrual cycle. This model leverages the current scientific understanding of the unimodal form of the relationship between the probability of becoming pregnant as a function of time across the fertile window. Furthermore, we analyzed previous studies on the effect of the fertile window day on the likelihood of pregnancy in order to propose a set of prior distributions.

Additionally, we propose missing data imputation mechanisms for sexual intercourse status, categorical covariates, and continuous covariates. Simulation studies and a real data analysis demonstrate the improved efficiency of these proposed models.

APPENDIX A: COMPOSITE QUANTILE-BASED CLASSIFIERS TECHNICAL DETAILS

A.1 Quantile classifier further discussion

A.1.1 Example differences of quantile distances

Two small examples are provided in Figure A.1 to further demonstrate the θ-th quantile classifier and the relationship between the quantile distances and the difference of the quantile distances. In these examples we suppose that $F_0^{-1}(\theta) < F_1^{-1}(\theta)$ at both the 0.25- and 0.75-th quantile levels, and that $F_1^{-1}(\theta) - F_0^{-1}(\theta)$ is the same for the two quantile levels. This is the case for example for populations that follow $N(0,1)$ and $N(1,1)$ distributions. In the upper-left plot we show the check loss distance to the populations' 0.25-th quantiles as a function of input z and denoted as Φ_1 and Φ_0, while in the bottom-left plot we show the difference of these same check loss distances, i.e. $\Lambda = \Phi_1 - \Phi_0$, again as a function of input z. The upper-right and lower right plots also show the check loss distances and difference between the distances, but instead with respect to the populations' 0.75-th quantiles. In each of these figures τ_θ can be seen to represent the decision rule boundary point, i.e. when $\Lambda = 0$. It can also be shown that the decision rule boundary is always between F_0^{-1} and F_1^{-1}, and that when the quantile level is less than 1/2 that τ_θ is closer to F_1^{-1}, and conversely that when the quantile level is greater than 1/2 that τ_θ is closer to F_0^{-1}. Finally, we note that Λ is piecewise linear with a slope of -1 between $F_0^{-1}(\theta)$ and $F_1^{-1}(\theta)$, and is constant otherwise.

A.1.2 Consistency of the optimal quantile classifier

We now present a result for the consistency of the empirical version that we make use of in Theorem 4 of the main paper. We note that this proposition is essentially a special case of Theorem 1 from Hennig and Viroli (2016). Let δ be a small positive constant, then we will need the following assumptions.

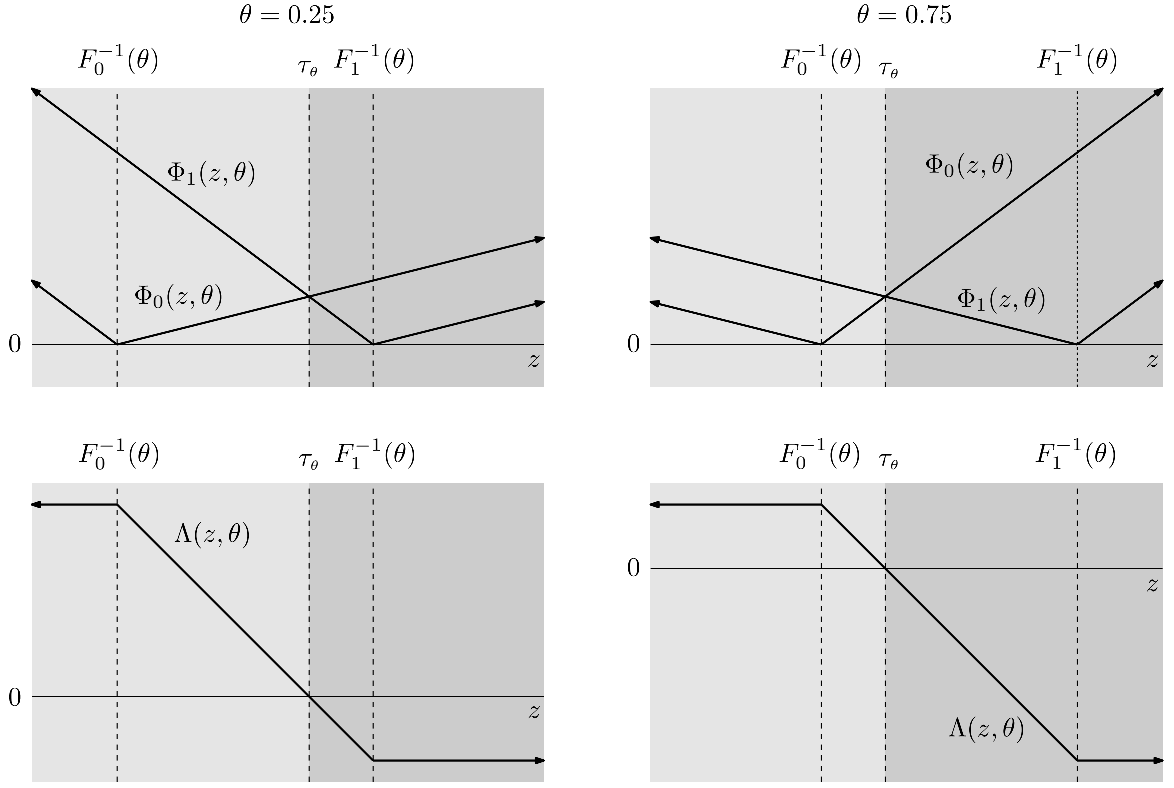
Assumption 1. F_i^{-1} is a continuous function of $\theta,\ i = 0,1$.

Assumption 2. $\mathbb{P}\left(\Lambda(Z,\theta) = 0\right) = 0$ for all $\theta \in [\delta, 1-\delta]$.

Assumption 3. There is a unique $\tilde{\theta}$ that satisfies $\tilde{\theta} = \arg\max_{\theta \in T} \Psi(\theta)$.

Proposition 6. *Let $\tilde{\theta}$ be a solution to $\tilde{\theta} = \arg\max_{\theta \in T} \Psi(\theta)$, and let $\hat{\theta}_n$ denote the empirically optimal quantile level. Then under Assumptions 1 and 2, it follows that $\Psi(\hat{\theta}_n) \xrightarrow{p} \Psi(\tilde{\theta})$. Furthermore, under Assumptions 1, 2, and 3, it follows that $\hat{\theta}_n \xrightarrow{p} \tilde{\theta}$.*

Figure A.1: Example within-class quantile distances and difference in quantile distances for two choices of quantiles.



It is worth noting that convergence to the optimal quantile level is based on the empirically optimal choice for a training data set without the need for additional training on a validation data set. Intuitively, this can be explained by the fact that the model complexity is the same for all of the possible quantile classifier models, so we do not need to include a model validation training step to defend against overfitting the model to the data.

A.1.3 Calculating the quantile classifier for a fixed quantile level

So far it has been established that the classification rate of the empirically optimal quantile classifier converges to the classification rate of the true optimal quantile classifier. In this section we investigate the practical considerations of calculating the empirically optimal quantile classifier. Recall the definition of the quantile classifier:

$$\text{For an observation } z\text{, classify to:} \quad \begin{cases} \Pi_0, & \text{if} \quad \Lambda\left(z, \theta\right) > 0, \\ \Pi_1, & \text{otherwise.} \end{cases}$$

Then an alternative version of the classifier is provided by Supplementary Proposition 7.

Proposition 7. *Assume for the moment that $F_0^{-1}(\theta) < F_1^{-1}(\theta)$. Then the quantile classifier defined in equation (2.1) can be equivalently expressed as follows.*

$$\textit{For an observation } z\textit{, classify to:} \quad \begin{cases} \Pi_0, & z \; < \; \theta\, F_0^{-1}(\theta) + (1-\theta)\, F_1^{-1}(\theta), \\ \Pi_1, & \text{otherwise.} \end{cases}$$

If on the other hand $F_0^{-1}(\theta) > F_1^{-1}(\theta)$, then the direction of the inequality changes so that we instead classify an observation to Π_0 when $z \; > \; \theta\, F_1^{-1}(\theta) + (1-\theta)\, F_0^{-1}(\theta)$, and to Π_1 otherwise.

It is interesting to note that the decision rule boundary lies on the line segment between $F_0^{-1}(\theta)$ and $F_1^{-1}(\theta)$, with the location of the point on the line segment determined by the quantile level. We saw an example of this in Figure A.1 in that for the 0.25-th quantile level the decision rule boundary was closer to the larger quantile, and for the 0.75-th quantile level the decision rule boundary was closer to the smaller quantile.

Next, the result given in Supplementary Proposition 8 provides an expression with which to obtain an optimal solution for the empirical quantile estimator. Recall the form of the θ-th empirical quantile:

$$\arg\min_q \left\{ \theta \sum_{x_i > q} |x_i - q| \; + \; (1-\theta) \sum_{x_i \le q} |x_i - q| \right\}. \tag{A.1}$$

We can view this optimization problem as a special case to the quantile regression problem introduced in Koenker and Bassett Jr (1978), where the x_i's correspond to the response variables, and q is a single regression coefficient corresponding to predictor data with only an intercept term and no other covariates. In light of this fact, we may utilize any of the theory developed in the quantile regression literature to solve this minimization problem. However, it turns out that there is a simple closed-form solution available for this particular special case of quantile regression, which is presented as Supplementary Proposition 8.

Proposition 8. *Let $x_1, \ldots, x_m$ be points on $\mathbb{R}$. Then a solution to equation (A.1) providing the θ-th empirical quantile for $x_1, \ldots, x_m$ is given by $\lceil m\theta \rceil$-th largest value of the x_i.*

Futhermore, as a consequence of the results of Supplementary Propositions 7 and 8, the result in Supplementary Proposition 9 characterizes the complexity of obtaining the empirical decision rule boundary for a given dataset.

Proposition 9. *Obtaining the quantile classifier decision rule boundary for a fixed choice of quantile level is an $\mathcal{O}(n)$ operation.*

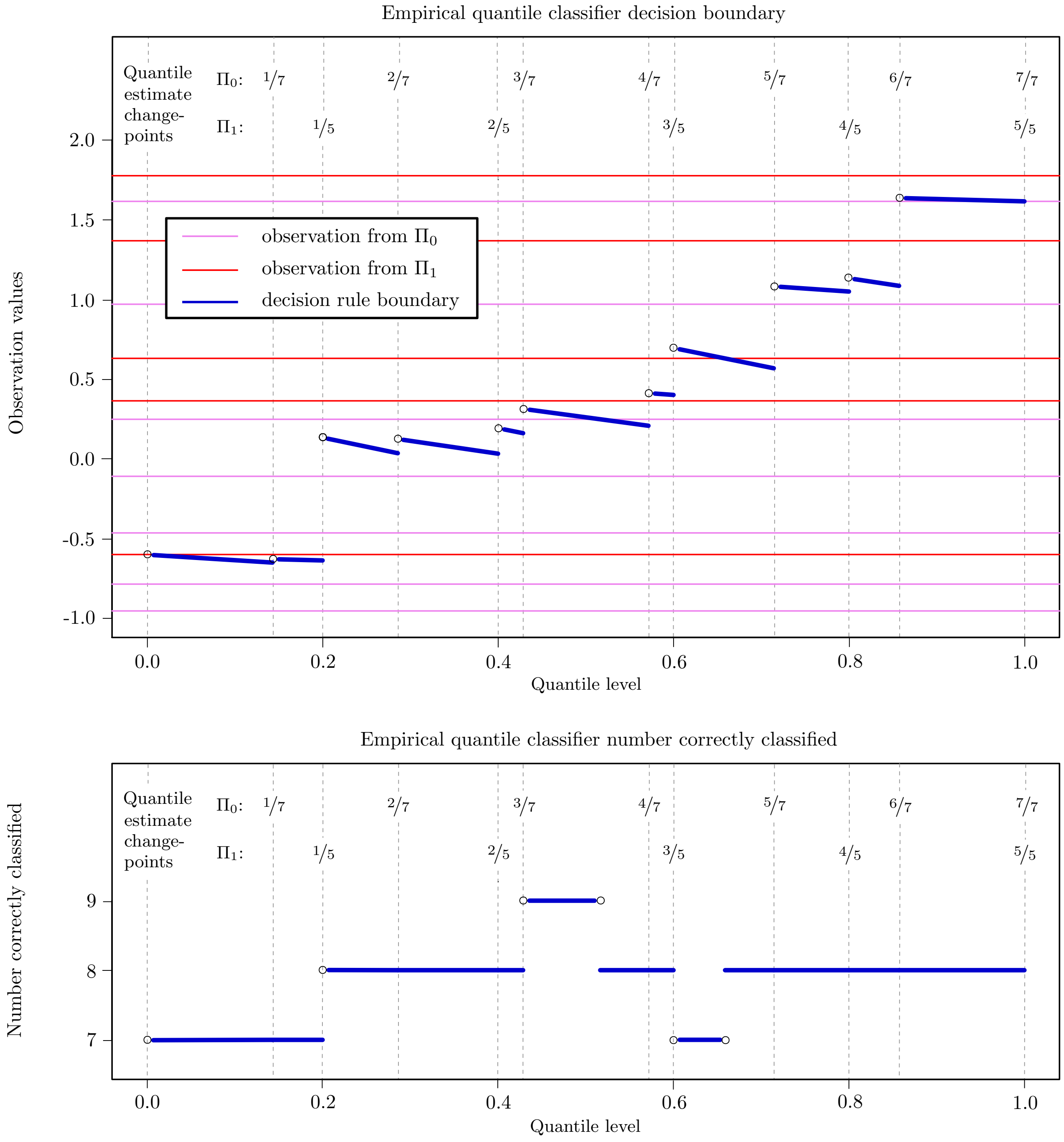
A.1.4 Calculating the empirically optimal quantile classifier

In the previous section we described how to calculate the quantile classifier for a fixed choice of quantile level. In this section we describe an algorithm used to construct the empirically optimal quantile classifier. To begin with, we start with an example calculation that illustrates the considerations that lead to the algorithm.

A.1.4.1 An example calculation

An illustrative example showing the decision rule boundary and corresponding classification rate as a function of the quantile level for a small data set is presented as Figure A.2. A data set for class 0 with seven observations was sampled from independent $N(0, 1)$ distributions yielding the values -0.96, $-0.78, -0.46, -0.10, 0.24, 0.98, 1.61$, and a data set for class 1 with five observations was sampled from independent $N(1, 1)$ distributions yielding the values $-0.60, 0.37, 0.64, 1.37, 1.78$. These values are shown as the horizontal lines in the top panel, along with the decision rule boundaries corresponding to this data which are displayed as the diagonal blue lines. The dashed vertical lines represent the quantile levels at which the quantile estimates change for at least one of the classes. In the bottom panel the number of correctly classified observations out of a possible 12 observations is plotted as a function of the quantile level. From this figure we see that the classification rate of the quantile classifier is constant for the intervals between the quantile estimate change points except when the set of corresponding decision rule boundary values includes values in the data. When this is the case the intervals can be further divided into sub-intervals with a constant classification rate. For this particular example the theoretically optimal quantile level is 0.5, which is indeed one of the empirically optimal quantile levels.

Figure A.2: Illustrative plot of the empirical classification rate for a toy example. An example data set is shown with 7 values from class Π_0 and 5 values from class Π_1.



From this figure, we can see that the decision rule boundaries are intervals as a function of the quantile level between levels at which the quantile estimates change for at least one of the classes (i.e. each of the blue line segments in the top panel of the figure). The algorithm described in the next section measures the classification rate for the points in each of intervals, yielding an exhaustive search of potential quantile levels.

A.1.4.2 Quantile classifier algorithm

In order to calculate the empirical classification rate as a function of quantile level, the first step is to obtain the set of valid decision rule boundaries. In other words, let $\phi(\theta;\ \text{data})\colon (0,1) \mapsto \mathbb{R}$ be the function that maps a quantile level for a given data set to the quantile classifier decision rule boundary; then we wish to obtain the set $\big\{x \in \mathbb{R}\colon\ \ \theta \in (0,1),\ \phi(\theta;\ \text{data}) = x\big\}$. The key observation that facilitates the calculation of this set is that the domain $(0,1)$ can be partitioned into smaller intervals for which range of the decision rule boundary function ϕ can be easily calculated. In more detail, consider an interval $(\theta_{\text{low}},\ \theta_{\text{high}}]$ such that $\lceil \theta n_0 \rceil$ and $\lceil \theta n_1 \rceil$ are each constant for all θ in the interval. As a consequence of the form of the quantile estimator, it follows that there are values a and b such that $F_{(0)}^{-1}(\theta) = a$ and $F_{(1)}^{-1}(\theta) = b$ for all θ in the interval. Thus by the result on the form of the decision rule boundary, we have $\phi(\theta;\ \text{data}) = \theta a + (1-\theta) b$ for any θ in the interval, and it follows that $\Big\{x \in \mathbb{R}\colon\ \ \theta \in (\theta_{\text{low}},\ \theta_{\text{high}}],\ \phi(\theta;\ \text{data}) = x\Big\} = \Big[\theta_{\text{high}}\, a + (1-\theta_{\text{high}})\, b,\quad \theta_{\text{low}}\, a + (1-\theta_{\text{low}})\, b\Big)$.

Next we consider how to partition $(0,1)$ into such sets. For notational convenience we actually consider $(0,1]$: even though the 1-th quantile doesn't make sense theoretically, its estimate is well-defined. Now $\lceil \theta n_0 \rceil$ is a step function with the endpoints for each step taking place at $1,\dots,n_0$ and which occur for values of θ at $1/n_0, 2/n_0, \dots, n_0/n_0$. Similarly, $\lceil \theta n_1 \rceil$ is a step function with the endpoint for each step taking place at $1,\dots,n_1$ and which occurs for values of θ at $1/n_1, 2/n_1, \dots, n_1/n_1$. Thus any interval that doesn't include one of those values of θ except as its upper bound is constant for each of these functions. So we can partition $(0,1]$ by letting each interval have an open lower bound be one of the step function change locations for one of the classes, and having the upper bound be the next smallest step function change location for either of the classes (and we also need an interval from 0 to the smallest step function change location).

Consider then one such interval, say $[x_{\text{low}}, x_{\text{high}})$, such that for each $x \in [x_{\text{low}}, x_{\text{high}})$ there exists a $\theta \in (0,1]$ with $\phi(\theta;\ \text{data}) = x$. At this point we would like to calculate the classification rate for all x in the interval: we can achieve this by conceptually "sliding" the decision rule boundary across the interval. Since the classification rate doesn't change as we slide the decision rule boundary in-between observations, we can find the values in the data that belong to $[x_{\text{low}}, x_{\text{high}})$ and partition the interval into sub-intervals that each have a constant classification rate. So for example, if $x_k, \ldots, x_{k+s}$ are the only observations in the data that belong to $[x_{\text{low}}, x_{\text{high}})$, then $[x_{\text{low}}, x_k),\ [x_k, x_{k+1}),\ \ldots,\ [x_{k+s-1}, x_{k+s}),\ [x_{k+s},\ x_{\text{high}})$ may be such a partitioning.

The reason we say that this 'may' be a partitioning is that whether the endpoints of the sub-intervals are open or closed depends on which class we have arbitrarily decided to classify an observation to in the event of ties. Note that when the quantile level θ corresponds to a decision rule boundary with a value equal to a point in the data, say x^*, then the quantile distance of that point to the populations' empirical θ-th quantiles is the same for either population, so we have to fall back on our classification rule in the event of ties. Whether the decision rule has the same classification rate as an open interval with upper bound x^* or as an open interval with lower bound x^* depends on which class has a smaller empirical θ-th quantile, and what our tiebreaker rule is. If the tiebreaker goes to the class with a larger empirical θ-th quantile, then the classification rate is the same as an open interval with with lower bound x^*, and we get a partitioning of $[x_{\text{low}}, x_{\text{high}})$ with the form $[x_{\text{low}}, x_k),\ [x_k, x_{k+1}),\ \ldots,\ [x_{k+s-1}, x_{k+s}),\ [x_{k+s},\ x_{\text{high}})$. If however, the tiebreaker goes to the class with a larger empirical θ-th quantile, then the classification rate is the same as an open interval with upper bound x^*, and we get a partitioning of $[x_{\text{low}}, x_{\text{high}})$ with the form $[x_{\text{low}}, x_k],\ (x_k, x_{k+1}],\ \ldots,\ (x_{k+s-1}, x_{k+s}],\ (x_{k+s},\ x_{\text{high}})$.

Finally, we note that we may wish to map back the sets from the decision rule boundary space to their quantile levels space, for example if we've found a set with a good empirical classification rate. Recall that we started with a set of quantile levels $(\theta_{\text{low}},\ \theta_{\text{high}}]$ such that F_0^{-1} and F_1^{-1} are each constant for all θ in the interval, and that each $x \in [x_{\text{low}}, x_{\text{high}})$ corresponds to the decision rule boundary for a θ from that interval. Let us write $F_{(0)}^{-1}(\theta) = a$ and $F_{(1)}^{-1}(\theta) = b$ for all θ in the interval, and assume that a is strictly less than b. Then for $x \in [x_{\text{low}}, x_{\text{high}})$, we observe the following relation:

$$x = \theta\, a + (1 - \theta)\, b \qquad \Longleftrightarrow \qquad \theta = \frac{b - x}{b - a}.$$

At this point we have all of the concepts that we need to construct an algorithm to calculate an empirically optimal choice of quantile level. The corresponding algorithm is presented as Algorithm 1, and a result for the algorithm complexity is presented as Supplementary Proposition 10.

Proposition 10. *The decision rule for the empirically optimal quantile classifier can be obtained in $\mathcal{O}(n^2)$ time.*

Algorithm 1: Calculating the empirically optimal quantile classifier

```
Data : $v_1, \dots, v_{n_0}$ from $\Pi_0$ and $w_1, \dots, w_{n_1}$ in $\Pi_1$
sort observations $v_1, \dots, v_{n_0}$ and $w_1, \dots, w_{n_1}$. Assume that in the case of ties, we classify to $\Pi_1$.
sort $v_1, \dots, v_{n_0}, w_1, \dots, w_{n_1}$ and denote the unique values as $x_1, \dots, x_r$
Produce a sorted set of quantile levels given by
  $\Theta = \Big\{0\Big\} \bigcup \Big\{\frac{k}{n_0} : 1 \le k \le n_0\Big\} \bigcup \Big\{\frac{k}{n_1} : 1 \le k \le n_1\Big\}$
for $i$ in 2 to $|\Theta|$ do
    calculate $\hat{F}^{-1}_{0n_0}(\theta_i)$ and $\hat{F}^{-1}_{1n_1}(\theta_i)$, and find
        $a = \min\Big\{\hat{F}^{-1}_{0n_0}(\theta_i),\ \hat{F}^{-1}_{1n_1}(\theta_i)\Big\}$    and    $b = \max\Big\{\hat{F}^{-1}_{0n_0}(\theta_i),\ \hat{F}^{-1}_{1n_1}(\theta_i)\Big\}$
    calculate the interval
        $G_i = \Big[\theta_i\, a + (1 - \theta_i)\, b,\quad \theta_{i-1}\, a + (1 - \theta_{i-1})\, b\Big) = \big[x_{\text{low}},\ x_{\text{high}}\big)$
    find the smallest $x_i \in [x_{\text{low}}, \infty)$. If $x_i \ge x_{\text{high}}$ then calculate the classification rate for $G_i$.
    while $x_i \in G_i$ do
        calculate the classification rate for the interval with upper and lower bounds determined by $x_i$
          and $\min\{x_{i+1}, x_{\text{high}}\}$, respectively (for whether bounds in interval are open or closed see
          below)
        if $\hat{F}^{-1}_{0n_0}(\theta_i) < \hat{F}^{-1}_{1n_1}(\theta_i)$ then
            calculate classification rate for the interval with bounds $x_i$ and $x_{i+1}$ among the following:
                $[x_{\text{low}}, x_k],\ (x_k, x_{k+1}],\ \dots,\ (x_{k+s-1}, x_{k+s}],\ (x_{k+s},\ x_{\text{high}})$
        else
            calculate classification rate for the interval with bounds $x_i$ and $x_{i+1}$ among the following:
                $[x_{\text{low}}, x_k),\ [x_k, x_{k+1}),\ \dots,\ [x_{k+s-1}, x_{k+s}),\ [x_{k+s},\ x_{\text{high}})$
        end
        map interval back to corresponding interval in the quantile levels space
        $x_i \leftarrow x_{i+1}$
    end
end
```

It remains an open question as to whether the bound obtained in Supplementary Proposition 10 is actually achievable. In practice, for both simulated and real data we typically see less than n total intervals that need to be considered in the outer and inner loops, so when this is the case the algorithm runs in $\mathcal{O}(n \log n)$ time. As an alternative, we can approximate the empirically optimal quantile classifier by simply calculating the empirical classification rate over a fine grid of quantile levels, which reduces the complexity to an $\mathcal{O}(Kn)$ operation where K is the number of points in the grid. This complexity is the case because each choice of quantile level requires calculating the decision rule boundary by finding the k_0-th and k_1-th largest value of each class for the appropriate values of k_0 and k_1, and then counting the number of observations on the correct side of the boundary, each of which is an $\mathcal{O}(n)$ operation.

One last issue that has yet to be addressed that is evident in this example is how to choose a quantile level based on a training data set when there is more than one empirically optimal value. When this is the case, as it typically is, we find that selecting the value closest to the median works well in many settings. One argument for this choice is that the median is the optimal choice of quantile level for two classes with symmetric distributions that differ only by a location shift. While this scenario may not hold in many settings, it still offers in some sense a conservative rule for what is hopefully a small set of empirically quantile levels from which to choose.

A.1.5 Calculating composite quantile-based classifiers

In the last section, we discussed calculation of the univariate quantile classifiers. In this section we consider the runtime for the calculation of the (multivariate) composite quantile-based quantile classifiers.

Two expressions for the run time complexity of Algorithm 2 from the main paper are presented in Supplementary Proposition 11. The first expression describes the complexity of a sequential implementation of the algorithm, while the second expression considers an implementation utilizing parallel computing. The opportunities for parallelism that are considered are the following. At the top level, the sub-models $f_1, \ldots, f_L$ can each be obtained independently of each other. Furthermore, within the calculation of each sub-model with index ℓ, the feature-specific calculations can each be performed independently of each other. Thus, for j in $1, \ldots p$, both the calculation of the optimal quantile level for the j-th feature with data $D_{\ell 1}$ and the transformation of the j-th component of every $\boldsymbol{x}_i$ in data $D_{\ell 2}$ can be calculated independently of the other features (lines 3-8 of Algorithm 2 from the main paper). Additionally, selection of the choice of linear combination coefficients can be parallelized over the folds in the cross-validation procedure (line 9 of Algorithm 2 from the main paper). Exploiting these opportunities for parallelism leads to the run time complexity described below.

Proposition 11. *Let T be the number of penalty levels and K be the number of folds considered when performing penalized logistic regression with k-fold cross-validation to select the linear combination coefficients $\alpha_0, \ldots, \alpha_p$. Then the model selection algorithm for composite quantile-based classifiers runs in $\mathcal{O}\Big(Lnp\big[n + KT\big]\Big)$ time. Furthermore, if the number of compute nodes in a computing cluster topology is at least L, and C denotes the smallest number of available cores from among the compute nodes, then the run time of Algorithm 2 from the main paper is reduced to $\mathcal{O}\Big(\frac{np}{C}\big[n + KT\big]\Big)$.*

A.2 Theoretical proofs

In this section we provide proofs for the propositions that were presented earlier in the supplementary materials (Supplementary Propositions 1-6), as well as a proof for Theorem 1-4 from the main paper. Additionally, in order to make Theorem 3 more readily digestible, we have split off a number of preliminary propositions that we make use of in the theorem. The proofs of Supplementary Propositions 1-6 are provided in Section A.2.1, the proofs for Theorems 1, 2, and 4 are provided in Section A.2.2, and the proof of Theorem 3 and the supporting propositions are provided in Section A.2.3.

A.2.1 Proofs of Supplementary Propositions 1-6

Proof of Supplementary Proposition 6. The fact that $\Psi(\hat{\theta}_n) \xrightarrow{p} \Psi(\tilde{\theta})$ is a special case of Theorem 1 in Hennig and Viroli (2016) for a feature-space of dimension 1. Furthermore, during the proof of that theorem it was shown that under Assumptions 1 and 2, Ψ is a continuous function of θ. To show that $\hat{\theta}_n \xrightarrow{p} \tilde{\theta}$ suppose that the claim doesn't hold. Then there exists an $\epsilon > 0$ and $\delta > 0$ such that for all $N \in \mathbb{N}$, there exists an $n \geq N$ such that

$$\mathbb{P}\left(\big|\hat{\theta}_n - \tilde{\theta}\big| > \epsilon\right) \geq \delta.$$

Now, because Ψ is continuous and $\Psi(\tilde{\theta})$ is a unique maximum, it follows that we can find $\nu > 0$ such that

$$\min\Big\{\big|\Psi(\tilde{\theta} - \epsilon) - \Psi(\tilde{\theta})\big|\,,\;\; \big|\Psi(\tilde{\theta} + \epsilon) - \Psi(\tilde{\theta})\big|\Big\} \geq \nu.$$

Therefore, for all $N \in \mathbb{N}$, there exists an $n \geq N$ such that

$$\mathbb{P}\left(\left|\Psi(\hat{\theta}_n) - \Psi(\tilde{\theta})\right| \geq \nu\right) \geq \mathbb{P}\left(\left|\hat{\theta}_n - \tilde{\theta}\right| \geq \epsilon\right) \geq \delta.$$

But this is in contradiction to the fact that $\Psi(\hat{\theta}_n) \xrightarrow{p} \Psi(\tilde{\theta})$. □

Proof of Supplementary Proposition 7. Suppose $F_{(0)}^{-1}(\theta) \neq F_{(1)}^{-1}(\theta)$. It is clear (e.g. see Figure A.1) that $\Phi_{(0)}(z, \theta)$ is equal to $\Phi_{(1)}(z, \theta)$ at exactly one point, say τ, and that the following holds:

$$\begin{cases} \Phi_{(0)}(z,\theta) < \Phi_{(1)}(z,\theta), & z < \tau, \\ \Phi_{(0)}(z,\theta) = \Phi_{(1)}(z,\theta), & z = \tau, \\ \Phi_{(0)}(z,\theta) > \Phi_{(1)}(z,\theta), & z > \tau. \end{cases}$$

Furthermore, we can infer that $F_{(0)}^{-1}(\theta) < \tau < F_{(1)}^{-1}(\theta)$. Setting the loss functions equal for z in this interval yields:

$$\begin{aligned}
&\Phi_{(0)}(z,\theta) \overset{set}{=} \Phi_{(1)}(z,\theta) \\
&\iff \quad \mathbb{1}\left(z > F_{(0)}^{-1}(\theta)\right) \theta \left(z - F_{(0)}^{-1}(\theta)\right) + \mathbb{1}\left(z \leq F_{(0)}^{-1}(\theta)\right) (1-\theta) \left(F_{(0)}^{-1}(\theta) - z\right) \\
&\qquad\qquad = \mathbb{1}\left(z > F_{(1)}^{-1}(\theta)\right) \theta \left(z - F_{(1)}^{-1}(\theta)\right) \\
&\qquad\qquad\qquad + \mathbb{1}\left(z \leq F_{(1)}^{-1}(\theta)\right) (1-\theta) \left(F_{(1)}^{-1}(\theta) - z\right) \\
&\iff \quad \theta \left(z - F_{(0)}^{-1}(\theta)\right) = (1-\theta) \left(F_{(1)}^{-1}(\theta) - z\right) \\
&\iff \quad z = \theta\, F_{(0)}^{-1}(\theta) + (1-\theta)\, F_{(1)}^{-1}(\theta).
\end{aligned}$$

It can then be verified that $\Phi_{(0)}(z,\theta) < \Phi_{(1)}(z,\theta)$ corresponds to classifying z to $\Pi_{(0)}$ and that $\Phi_{(0)}(z,\theta) > \Phi_{(1)}(z,\theta)$ corresponds to classifying z to $\Pi_{(1)}$. Combining these facts yields the desired result. □

We can write the minimization problem,

$$\min_q \left\{ \theta \sum_{x_i > q} |x_i - q| \; + \; (1-\theta) \sum_{x_i \leq q} |x_i - q| \right\}$$

equivalently as

$$\begin{aligned} \underset{q^+,q^-,\boldsymbol{u},\boldsymbol{v}}{\text{minimize}} \quad & \theta \textstyle\sum_{i=1}^m u_i \;+\; (1-\theta) \textstyle\sum_{i=1}^m v_i \\ \text{subject to} \quad & x_i - (q^+ - q^-) = u_i - v_i, \quad i = 1, \dots, m \\ & q^+ \geq 0,\ q^- \geq 0,\ \boldsymbol{u} \geq \boldsymbol{0},\ \boldsymbol{v} \geq \boldsymbol{0} \end{aligned}$$

which is seen to be a linear programming problem in standard form. By rewriting the equality condition as $q^+ - q^- + u_i - v_i = x_i$ for $i = 1, \dots, m$, we can express the equality condition in matrix form as

$$\begin{bmatrix} 1 & -1 & 1 & & & -1 & & \\ \vdots & \vdots & & \ddots & & & \ddots & \\ 1 & -1 & & & 1 & & & -1 \end{bmatrix} \begin{bmatrix} q^+ \\ q^- \\ \boldsymbol{u} \\ \boldsymbol{v} \end{bmatrix} = \begin{bmatrix} x_1 \\ \vdots \\ x_m \end{bmatrix}.$$

Recall that a solution for $(q^+, q^-, \boldsymbol{u}, \boldsymbol{v})$ is a basic solution if and only if there exist indices $B(1), \dots, B(m)$ such that both (i) the columns of the coefficient matrix with column indices subset by $B(1), \dots, B(m)$ are linearly independent, and (ii) if an element of $(q^+, q^-, \boldsymbol{u}, \boldsymbol{v})$ does not correspond to one of $B(1), \dots, B(m)$ then the element must have a value of 0.

We can see that in order to have independent columns from the coefficient matrix, then for each i, no more than one of the columns corresponding to either u_i or v_i can have an index in $B(1), \dots, B(m)$, and additionally no more than one of the columns corresponding to either q^+ or q^- can have an index in $B(1), \dots, B(m)$. Furthermore, we note that if we have one column corresponding to q^+ or q^-, and one column corresponding to either u_i or v_i for each i, then we have $m + 1$ columns, which is still one column too many. Thus we see that there must be either (i) exactly one column corresponding to either u_i or v_i for each i and no columns corresponding to either q^+ or q^-, or (ii) exactly one column corresponding to either u_i or v_i for each i less one and exactly one column corresponding to either q^+ or q^-.

Furthermore, we can infer that if a column corresponding to q^+ or q^- has an index in $B(1), \dots, B(m)$ and the solution is feasible (i.e. q^+ and q^- are both nonnegative), then $q^+ = (x_i)_+$ and $q^- = (-x_i)_+$, where the index i corresponds to the only i without a column corresponding to either u_i or v_i, and $(z)_+ = \max(0, z)$. Let $q = q^+ - q^-$, then it follows that a feasible solution for any $j \neq i$ has $u_i = (x_i - q)_+$ and $v_i = (q - x_i)_+$.

This leads to a set of basic feasible solutions given by $q \in \{x_1, \ldots, x_m\}$. One last basic feasible solution is given for $q = 0$ with $u_i = (x_i)_+$ and $v_i = (-x_i)_+$ for all i.

Next we aim to find the minimizing basic feasible solution. Suppose that $\lceil \theta m \rceil = k$ and that $\ell < k$, and let $q = x_k$ and $q' = x_\ell$. Then comparing the objective function evaluated at q and q', we have

$$\begin{aligned}
&\left\{\theta \sum_{i=\ell+1}^{m} (x_i - q') + (1-\theta)\sum_{i=1}^{\ell}(q' - x_i)\right\} - \left\{\theta \sum_{i=k+1}^{m} (x_i - q) + (1-\theta)\sum_{i=1}^{k}(q - x_i)\right\} \\
&\qquad = (1-\theta)\sum_{i=1}^{\ell}\Big\{(q' - x_i) - (q - x_i)\Big\} \\
&\qquad\quad + \theta \sum_{i=\ell+1}^{k} (x_i - q') - (1-\theta) \sum_{i=\ell+1}^{k} (q - x_i) \\
&\qquad\quad + \theta \sum_{i=k+1}^{m} \Big\{(x_i - q') - (x_i - q)\Big\} \\
&\qquad = (1-\theta)\sum_{i=1}^{\ell}\Big\{(q' - x_i) - (q - x_i)\Big\} \\
&\qquad\quad + \theta \sum_{i=\ell+1}^{k} (x_i - q') - (1-\theta) \sum_{i=\ell+1}^{k} (q - x_i) \pm (1-\theta) \sum_{i=\ell+1}^{k} (q' - x_i) \\
&\qquad\quad + \theta \sum_{i=k+1}^{m} \Big\{(x_i - q') - (x_i - q)\Big\} \\
&\qquad = -(1-\theta)\sum_{i=1}^{\ell} (q - q') \\
&\qquad\quad + \theta \sum_{i=\ell+1}^{k} (x_i - q') - (1-\theta) \sum_{i=\ell+1}^{k} (q - q') - (1-\theta) \sum_{i=\ell+1}^{k} (q' - x_i) \\
&\qquad\quad + \theta \sum_{i=k+1}^{m} (q - q') \\
&\qquad = -(1-\theta)\sum_{i=1}^{k} (q - q') \\
&\qquad\quad + \theta \sum_{i=\ell+1}^{k} (x_i - q') - (1-\theta) \sum_{i=\ell+1}^{k} (q' - x_i) \\
&\qquad\quad + \theta \sum_{i=k+1}^{m} (q - q')
\end{aligned}$$

$$
\begin{aligned}
&= -(1-\theta)\sum_{i=1}^{k}(q-q') \\
&\qquad + \sum_{i=\ell+1}^{k}(x_i-q') \\
&\qquad + \theta\sum_{i=k+1}^{m}(q-q') \\
&= -(1-\theta)\,k\,(q-q') + \sum_{i=\ell+1}^{k}(x_i-q') + \theta(m-k)(q-q') \\
&= \sum_{i=\ell+1}^{k}(x_i-q') - (k-\theta m)(q-q') \\
&\geq (x_k-q') - (q-q') \\
&= x_k - q \\
&= 0.
\end{aligned}
$$

The inequality is due to the fact that $x_i \geq q'$ for $i \geq \ell+1$, and also the fact that $k-\theta m < 1$. Suppose now that $\lceil \theta m \rceil = k$ and that $\ell > k$, and let $q = x_k$ and $q' = x_\ell$. Then comparing the objective function evaluated at q and q', we have

$$
\begin{aligned}
&\left\{\theta\sum_{i=\ell+1}^{m}(x_i-q') + (1-\theta)\sum_{i=1}^{\ell}(q'-x_i)\right\} - \left\{\theta\sum_{i=k+1}^{m}(x_i-q) + (1-\theta)\sum_{i=1}^{k}(q-x_i)\right\} \\
&= (1-\theta)\sum_{i=1}^{k}\Big\{(q'-x_i)-(q-x_i)\Big\} \\
&\qquad + \theta\sum_{i=k+1}^{\ell}(x_i-q') - (1-\theta)\sum_{i=k+1}^{\ell}(q-x_i) \\
&\qquad + \theta\sum_{i=\ell+1}^{m}\Big\{(x_i-q')-(x_i-q)\Big\} \\
&= (1-\theta)\sum_{i=1}^{k}(q'-q) \\
&\qquad - \theta\sum_{i=k+1}^{\ell}(q'-q) + \sum_{i=k+1}^{\ell}(x_i-q) \\
&\qquad - \theta\sum_{i=l+1}^{m}(q'-q)
\end{aligned}
$$

$$
\begin{aligned}
&= (1-\theta)\sum_{i=1}^{k}(q'-q) + \sum_{i=k+1}^{\ell}(x_i-q) - \theta\sum_{i=k+1}^{m}(q'-q) \\
&= (1-\theta)\,k\,(q'-q) + \sum_{i=k+1}^{\ell}(x_i-q) - \theta\,(m-k)(q'-q) \\
&= (k-\theta m)(q'-q) + \sum_{i=k+1}^{\ell}(x_i-q) \\
&\geq 0.
\end{aligned}
$$

There is one basic feasible solution remaining to check, that where $q = 0$. To show that this is not the optimal solution except in the case that $x_{\lceil \theta m \rceil} = 0$, we make the following argument. Note that if we add some nonzero value τ to each x_i and if q^* is an optimal choice for the original problem, then $q^* + \tau$ is an optimal choice for the new problem. Choose $-\tau$ to be one of the x_i for some i: then 0 is one of the values in the transformed data, which was shown in the previous results to be no better than the $\lceil \theta m \rceil$-th largest value of the transformed data, so it follows that the $\lceil \theta m \rceil$-th largest value is optimal. Since τ cannot be better than optimal for the transormed data, then by the law of the contrapositive 0 cannot be better than optimal for the original data.

Proof of Supplementary Proposition 9. Constructing the θ-th quantile classifier requires merely finding the $\lceil \theta n_0 \rceil$-th largest value from the observations that were drawn from population Π_0, and the $\lceil \theta n_1 \rceil$-th largest value from the observations that were drawn from population Π_1, and then calculating the decision boundary based on Supplementary Proposition 7. Finding the k-th largest value of a set is an $\mathcal{O}(n)$ operation, and the calculation to obtain the decision boundary using Supplementary Proposition 7 is an $\mathcal{O}(1)$ operation. □

Proof of Supplementary Proposition 10. Sorting the data is an $\mathcal{O}(n \log n)$ operation. The operations performed in line 2 of Algorithm 1 takes $\mathcal{O}(n)$ time.

The outer loop beginning on line 3 of Algorithm 1 requires some number of iterations that is bounded from above by $n - 1$. Within each iteration, calculating F_V^{-1} and F_W^{-1} and the interval G_i are each constant time operations.

The step in line 6 of Algorithm 1 is really a high-level view of a second loop. This inner loop requires first finding the set of x_i's with values in $\left[x_{\text{low}},\, x_{\text{high}}\right)$ which is an $\mathcal{O}(\log n)$ operation. The number of times that the inner loop is performed is determined by the number of points with values in the interval and is bounded from above by $n-1$. Calculating the classification rate the sub-interval in each step of the inner loop is a constant time operation, as is mapping the sub-interval back to the quantile levels space.

So combining the worst-case bound for the outer and inner loops, we find that the total number of intervals for which we calculate the classification rate for has a worst-case bound on the order of n^2 intervals, and that each calculation is a constant time operation for a total cost of $\mathcal{O}(n^2)$ time. The initial operations performed in lines 1-2 of Algorithm 1 have an aggregate cost with $\mathcal{O}(n \log n)$ time, so in total the algorithm runs in $\mathcal{O}(n^2)$ time. □

Proof of Supplementary Proposition 11. Consider the operations performed inside the outer-level for loop spanning lines 1-10 in Algorithm 2 from the main paper. Splitting the data into two parts is an $\mathcal{O}(n)$ operation. Next, consider the second-level for loop iterating over the features. We saw in Supplementary Proposition 10 that the decision rule for the empirically optimal quantile classifier for the j-th feature can be obtained in $\mathcal{O}(n^2)$ time. Next, calculating x^*_{ij} is a constant-time operation for an $\mathcal{O}(n)$ number of calculations, which in total is an $\mathcal{O}(n)$ operation.

Next, we break down the steps required in line 9 of Algorithm 2 from the main paper to select the linear combination coefficients via penalized logistic regression. Using the coordinate descent algorithm proposed in Friedman et al. (2007, 2010) yields an $\mathcal{O}(np)$ run time for a fixed choice of penalty parameter. Thus for a grid of size T penalty parameters and performing k-fold cross-validation for a total of K folds yields a total run time bound of $\mathcal{O}(KTnp)$.

Then, since the outer loop is performed a total of L times, we obtain the following run time bound for a sequential implementation of Algorithm 2:

$$\mathcal{O}\Big(L\Big[n + p(n^2+n) + KTnp\Big]\Big) = \mathcal{O}\Big(Lnp\big[n + KT\big]\Big).$$

Next, if we have L compute nodes available, we can assign the calculation of each sub-model f_i to one of the nodes; this reduces the problem to calculating the run time for each of the f_i. If we perform the parallelism over first the features (i.e. lines 3-8 of Algorithm 2 from the main paper), and then over the cross-validation folds (i.e. line 9 of Algorithm 2 from the main paper), then the run time bound for an implementation of Algorithm 2 utilizing parallel computing is given by

$$\mathcal{O}\left(n + \frac{p}{C}(n^2+n) + \frac{K}{C}Tnp\right) = \mathcal{O}\Big(\frac{np}{C}\big[n + KT\big]\Big).$$

This calculation shows that the cost of training the algorithm has a polynomial-time complexity. □

A.2.2 Proof of Theorems 1, 2, and 4

Proof of Theorem 1. The proof is largely adapted from Theorem 2 of Hennig and Viroli (2016). Suppose that the claim doesn't hold. Then there exists an $\epsilon > 0$, a $\delta > 0$, a sequence M of positive integers, and a sequence of quantile levels $\{\theta^*_m\}_{m \in M}$ such that

$$\mathbb{P}\left(\left|\Psi^\dagger_m(\theta^*_m) - \Psi(\theta^*_m)\right| > \epsilon\right) \geq \delta \quad \text{for all } m. \tag{A.2}$$

Now, every bounded sequence has at least one convergent subsequence, and we will choose to restrict our attention to such a subsequence in what follows. Define $\theta^* = \lim_{m\to\infty} \theta^*_m$, then we have

$$\begin{aligned}
&\mathbb{P}\left(\left|\Psi^\dagger_m(\theta^*_m) - \Psi(\theta^*_m)\right| > \epsilon\right)\\
&\quad= \mathbb{P}\left(\left|\Psi^\dagger_m(\theta^*_m) - \Psi^\dagger_m(\theta^*) + \Psi^\dagger_m(\theta^*) - \Psi(\theta^*) + \Psi(\theta^*) - \Psi(\theta^*_m)\right| > \epsilon\right)\\
&\quad\leq \mathbb{P}\left(\left|\Psi^\dagger_m(\theta^*_m) - \Psi^\dagger_m(\theta^*)\right| + \left|\Psi^\dagger_m(\theta^*) - \Psi(\theta^*)\right| + \left|\Psi(\theta^*) - \Psi(\theta^*_m)\right| > \epsilon\right)\\
&\quad\leq \mathbb{P}\left(\left|\Psi^\dagger_m(\theta^*_m) - \Psi^\dagger_m(\theta^*)\right| + \left|\Psi(\theta^*) - \Psi(\theta^*_m)\right| > \frac{\epsilon}{2}\right) \qquad\qquad (A.3)\\
&\qquad\quad + \mathbb{P}\left(\left|\Psi^\dagger_m(\theta^*) - \Psi(\theta^*)\right| > \frac{\epsilon}{2}\right).
\end{aligned}$$

It was established in Hennig and Viroli (2016) that Ψ is continuous under Assumptions 1 and 2, so as a result it follows that $\left|\Psi(\theta^*) - \Psi(\theta^*_m)\right| \to 0$. Now,

$$\begin{aligned}
&\Psi^\dagger_m(\theta^*_m) - \Psi^\dagger_m(\theta^*)\\
&\quad= \pi_0 \int \mathbb{1}\left(\Lambda_m(z, \theta^*_m) > 0\right) dP_0(z) + \pi_1 \int \mathbb{1}\left(\Lambda_m(z, \theta^*_m) \leq 0\right) dP_1(z)\\
&\qquad\quad - \pi_0 \int \mathbb{1}\left(\Lambda_m(z, \theta^*) > 0\right) dP_0(z) - \pi_1 \int \mathbb{1}\left(\Lambda_m(z, \theta^*) \leq 0\right) dP_1(z).\\
&\quad= \pi_0 \int \left[\mathbb{1}\left(\Lambda_m(z, \theta^*_m) > 0\right) - \mathbb{1}\left(\Lambda_m(z, \theta^*) > 0\right)\right] dP_0(z) \qquad\qquad (A.4)\\
&\qquad\quad + \pi_1 \int \left[\mathbb{1}\left(\Lambda_m(z, \theta^*_m) \leq 0\right) - \mathbb{1}\left(\Lambda_m(z, \theta^*) \leq 0\right)\right] dP_1(z).
\end{aligned}$$

From the proof of Theorem 1 in Hennig and Viroli (2016), we have that $\left|\hat{F}^{-1}_{im}(\theta^*_m) - \hat{F}^{-1}_{im}(\theta^*)\right| \xrightarrow{a.s.} 0,\ i = 0, 1$, and since the sample quantiles are strongly consistent, it follows that $\hat{F}^{-1}_{im}(\theta^*_m) \xrightarrow{a.s.} F^{-1}_i(\theta^*),\ i = 0, 1.$

Let us denote $\tau_{\theta^*_m} = \theta^*_m \hat{F}^{-1}_{(0),m}(\theta^*_m) + (1-\theta^*_m)\hat{F}^{-1}_{(1),m}(\theta^*_m)$, and $\tau_{\theta^*} = \theta^* \hat{F}^{-1}_{(0),m}(\theta^*) + (1-\theta^*)\hat{F}^{-1}_{(1),m}(\theta^*)$, then we also obtain that $|\tau_{\theta^*} - \tau_{\theta^*_m}| \xrightarrow{a.s.} 0$.

From Theorem 3 in Mason (1982), and under Assumption 1, we get $\lim_{m\to\infty}\sup_{\theta\in T}|\hat{F}^{-1}_{im}(\theta) - F^{-1}_i(\theta)|$ $= 0$ almost surely. Suppose for the time-being that $F^{-1}_0(\theta^*) < F^{-1}_1(\theta^*)$. Then for large values of m we will observe both $\hat{F}^{-1}_{0m}(\theta^*_m) < \hat{F}^{-1}_{1m}(\theta^*_m)$ and $\hat{F}^{-1}_{0m}(\theta^*) < \hat{F}^{-1}_{1m}(\theta^*)$ with probability 1. When this is the case then equation (A.4) can be expressed as

$$\begin{aligned}
\pi_0 \int & \Big[\mathbb{1}\Big(\Lambda_m(z,\theta^*_m) > 0\Big) - \mathbb{1}\Big(\Lambda_m(z,\theta^*) > 0\Big)\Big]\, dP_0(z) \\
&\quad + \pi_1 \int \Big[\mathbb{1}\Big(\Lambda_m(z,\theta^*_m) \le 0\Big) - \mathbb{1}\Big(\Lambda_m(z,\theta^*) \le 0\Big)\Big]\, dP_1(z) \\
&= \pi_0 \int \Big[\mathbb{1}\left(z < \tau_{\theta^*_m}\right) - \mathbb{1}\left(z < \tau_{\theta^*}\right)\Big]\, dP_0(z) \\
&\quad + \pi_1 \int \Big[\mathbb{1}\left(z \ge \tau_{\theta^*_m}\right) - \mathbb{1}\left(z \ge \tau_{\theta^*}\right)\Big]\, dP_1(z) \\
&= \pi_0 \int_{\tau_{\theta^*_m}}^{\tau_{\theta^*}} dP_0(z) + \pi_1 \int_{\tau_{\theta^*}}^{\tau_{\theta^*_m}} dP_1(z). \qquad \text{(A.5)}
\end{aligned}$$

On the other hand, if $F^{-1}_0(\theta^*) > F^{-1}_1(\theta^*)$, then for large values of m we will observe both $\hat{F}^{-1}_{0m}(\theta^*_m) > \hat{F}^{-1}_{1m}(\theta^*_m)$ and $\hat{F}^{-1}_{0m}(\theta^*) > \hat{F}^{-1}_{1m}(\theta^*)$ with probability 1. When this is the case then equation (A.4) can be expressed as

$$\begin{aligned}
\pi_0 \int & \Big[\mathbb{1}\Big(\Lambda_m(z,\theta^*_m) > 0\Big) - \mathbb{1}\Big(\Lambda_m(z,\theta^*) > 0\Big)\Big]\, dP_0(z) \\
&\quad + \pi_1 \int \Big[\mathbb{1}\Big(\Lambda_m(z,\theta^*_m) \le 0\Big) - \mathbb{1}\Big(\Lambda_m(z,\theta^*) \le 0\Big)\Big]\, dP_1(z) \\
&= \pi_0 \int \Big[\mathbb{1}\left(z > \tau_{\theta^*_m}\right) - \mathbb{1}\left(z > \tau_{\theta^*}\right)\Big]\, dP_0(z) \\
&\quad + \pi_1 \int \Big[\mathbb{1}\left(z \le \tau_{\theta^*_m}\right) - \mathbb{1}\left(z \le \tau_{\theta^*}\right)\Big]\, dP_1(z) \\
&= \pi_0 \int_{\tau_{\theta^*}}^{\tau_{\theta^*_m}} dP_0(z) + \pi_1 \int_{\tau_{\theta^*_m}}^{\tau_{\theta^*}} dP_1(z). \qquad \text{(A.6)}
\end{aligned}$$

But since $|\tau_{\theta^*} - \tau_{\theta^*_m}| \xrightarrow{a.s.} 0$, we observe that both (A.5) and (A.6) converge to 0, which establishes the almost sure convergence of the first term in (A.3).

Next, we consider the second term from (A.3). If we can show that $\Psi^{\dagger}_m(\theta^*)$ is continuous in $\big(F^{-1}_{0m}(\theta^*), F^{-1}_{1m}(\theta^*)\big)$, then we can invoke the continuous mapping theorem to establish convergence of $\Psi^{\dagger}_m(\theta^*)$ to $\Psi(\theta^*)$. To establish continuity, we look to find the limit of $\Psi^{\dagger}_m(\theta^*_m)$.

Consider $\big(F^{-1}_{0m}(\theta^*), F^{-1}_{1m}(\theta^*)\big)$ for some $\theta^* \in T$, then $\mathbb{1}\left(\Lambda_m(z,\theta^*) > 0\right)$ is equal to either $\mathbb{1}\left(z < \tau_{\theta^*_m}\right)$ or $\mathbb{1}\left(z > \tau_{\theta^*_m}\right)$, depending on whether $F^{-1}_{0m}(\theta^*) < F^{-1}_{1m}(\theta^*)$. It follows that for fixed $z \neq \tau_{\theta^*_m}$, then $\mathbb{1}\left(\Lambda_m(z,\theta^*) > 0\right)$ is continuous in $\big(F^{-1}_0(\theta^*), F^{-1}_1(\theta^*)\big)$. A nearly identical argument establishes continuity for $\mathbb{1}\left(\Lambda_m(z,\theta^*) \leq 0\right)$ on the same set of z.

Then, if we take the limit of $\Psi^{\dagger}_m(\theta^*_m)$, we can invoke the dominated convergence theorem to move the limit sign inside each of the integrals. Furthermore, since we have shown that the indicator functions are continuous except on a set with probability 0, we find that the limit of $\Psi^{\dagger}_m$ is the same as when evaluated at the limit of the sample quantiles, establishing continuity of $\Psi^{\dagger}_m(\theta^*)$.

Since the sample quantiles are strongly consistent, we can then use the continuous mapping theorem to establish that $\Psi^{\dagger}_m(\theta^*) \xrightarrow{a.s.} \Psi(\theta^*)$, from which it follows that $\left|\Psi^{\dagger}_m(\theta^*) - \Psi(\theta^*)\right| \xrightarrow{a.s.} 0$, which in turn implies convergence in probability of the second term in (A.3). This and our earlier results force both terms to 0 in the limit, however this is a contradiction to the statement in (A.2), so we conclude that the claim in Theorem 1 holds. □

Proof of Theorem 2. Let us consider the subproblem for the region given by the interval (x_k, x_{k+1}). Let M be a random variable that takes a value of 0 if Z is drawn from Π_0 and a value of 1 if Z is drawn from Π_1. Then

$$f_i\big(z \,|\, z \in (x_k, x_{k+1})\big) = \frac{f_i(z)}{F_i(x_{k+1}) - F_i(x_k)}, \qquad i = 0, 1, \tag{A.7}$$

and furthermore

$$\begin{aligned} &\mathbb{P}\Big(M = i \,\Big|\, Z \in (x_k, x_{k+1})\Big) \\ &\quad= \frac{\mathbb{P}\Big(Z \in (x_k, x_{k+1}) \,\Big|\, M = i\Big)\,\mathbb{P}\left(M = i\right)}{\mathbb{P}\Big(Z \in (x_k, x_{k+1})\Big)} \\ &\quad= \frac{\mathbb{P}\Big(Z \in (x_k, x_{k+1}) \,\Big|\, M = i\Big)\,\mathbb{P}\left(M = i\right)}{\sum_{i=0}^{1}\mathbb{P}\Big(Z \in (x_k, x_{k+1}) \,\Big|\, M = i\Big)\,\mathbb{P}\left(M = i\right)}, \end{aligned} \tag{A.8}$$

$$= \frac{\pi_i \left(F_i(x_{k+1}) - F_i(x_k) \right)}{\sum_{i=0}^{1} \pi_i \left(F_i(x_{k+1}) - F_i(x_k) \right)}, \qquad i = 0, 1.$$

Let us denote the density function defined in (A.7) as $f_i^{(k)}$, and let us denote the quantity defined in (A.8) as $\pi_i^{(k)}$, each for $\ i = 0, 1,\ k = 1, \ldots, t$. Then

$$\begin{aligned} \pi_i^{(k)} f_i^{(k)}(x) &= \frac{\pi_i \left(F_i(x_{k+1}) - F_i(x_k) \right)}{\sum_{i=0}^{1} \pi_i \left(F_i(x_{k+1}) - F_i(x_k) \right)} \, \frac{f_i(z)}{F_i(x_{k+1}) - F_i(x_k)} \\ &= \frac{\pi_i \, f_i(z)}{\sum_{i=0}^{1} \pi_i \left(F_i(x_{k+1}) - F_i(x_k) \right)}. \end{aligned} \tag{A.9}$$

Thus we see that $\pi_0 f_0(z)$ and $\pi_1 f_1(z)$ are multiplied by the same positive constant to obtain $\pi_0^{(k)} f_0^{(k)}(z)$ and $\pi_1^{(k)} f_1^{(k)}(z)$, respectively. Since it is assumed that there is exactly one value z_k^* in $(x_k,\, x_{k+1})$ such that $\pi_0 f_0(z_k^*) = \pi_1 f_1(z_k^*)$, it follows from (A.9) that z_k^* is also the only point such that $\pi_0^{(k)} f_0^{(k)}(z_k^*) = \pi_1^{(k)} f_1^{(k)}(z_k^*)$. Furthermore, for any points z such that $\pi_0 f_0(z) < \pi_1 f_1(z)$, it follows that $\pi_0^{(k)} f_0^{(k)}(z) < \pi_1^{(k)} f_1^{(k)}(z)$, and for any points z such that $\pi_0 f_0(z) > \pi_1 f_1(z)$, it follows that $\pi_0^{(k)} f_0^{(k)}(z) > \pi_1^{(k)} f_1^{(k)}(z)$. Thus the conditions of Lemma 2 in the supplementary materials of Hennig and Viroli (2016) are satisfied, and the quantile classifier achieves the Bayes rule classification rate for the region. Since this is the case for all of the regions, we conclude that the Bayes rule classification rate is achieved for the entire domain due to the form of the multimodal quantile classifier as defined in (11). □

Proof of Theorem 4. The upper bound on the difference between the component-wise quantile distances shown in (A.10) was established in the proof of Theorem 1 in Hennig and Viroli (2016). For $z \in \mathbb{R}$ and $\theta, \theta' \in (0, 1)$ then

$$\left| \Phi_{ij}(z, \theta) - \Phi_{ij}(z, \theta') \right| \leq |z| \, |\theta - \theta'| + 4 |F_{ij}^{-1}(\theta) - F_{ij}^{-1}(\theta')|, \qquad i = 1, 2,\ j = 1, \ldots, p. \tag{A.10}$$

Since F_{ij}^{-1} is continuous by assumption, it follows that for arbitrary fixed z, $\Phi_{ij}(z, \theta)$ is continuous in θ for every $\{i, j\}$. This in turn implies that for arbitrary fixed $\boldsymbol{z}$ then $\Lambda(\boldsymbol{z}, \boldsymbol{\theta}) = \alpha_0 + \sum_{j=1}^{p} \alpha_j \Lambda_j(z_j, \theta_j)$ is also continuous in $\boldsymbol{\theta}$, since Λ is a linear combination of the $\Phi_{ij}(z_j, \theta_j)$'s. Then we observe that

$$\lim_{\boldsymbol{\theta} \to \boldsymbol{\theta}^*} \Psi(\boldsymbol{\theta}) = \lim_{\boldsymbol{\theta} \to \boldsymbol{\theta}^*} \left\{ \pi_0 \int \mathbb{1} \left(\Lambda(\boldsymbol{z}, \boldsymbol{\theta}) > 0 \right) dP_0(\boldsymbol{z}) + \pi_1 \int \mathbb{1} \left(\Lambda(\boldsymbol{z}, \boldsymbol{\theta}) \leq 0 \right) dP_1(\boldsymbol{z}) \right\} \tag{A.11}$$

$$
\begin{aligned}
&= \pi_0 \int \lim_{\boldsymbol{\theta}\to\boldsymbol{\theta}^*} \mathbb{1}\left(\Lambda(\boldsymbol{z},\boldsymbol{\theta}) > 0\right) dP_0(\boldsymbol{z}) + \pi_1 \int \lim_{\boldsymbol{\theta}\to\boldsymbol{\theta}^*} \mathbb{1}\left(\Lambda(\boldsymbol{z},\boldsymbol{\theta}) \le 0\right) dP_1(\boldsymbol{z}) \\
&= \pi_0 \int \mathbb{1}\left(\lim_{\boldsymbol{\theta}\to\boldsymbol{\theta}^*} \Lambda(\boldsymbol{z},\boldsymbol{\theta}) > 0\right) dP_0(\boldsymbol{z}) + \pi_1 \int \mathbb{1}\left(\lim_{\boldsymbol{\theta}\to\boldsymbol{\theta}^*} \Lambda(\boldsymbol{z},\boldsymbol{\theta}) \le 0\right) dP_1(\boldsymbol{z}) \\
&= \pi_0 \int \mathbb{1}\left(\Lambda(\boldsymbol{z},\boldsymbol{\theta}^*) > 0\right) dP_0(\boldsymbol{z}) + \pi_1 \int \mathbb{1}\left(\Lambda(\boldsymbol{z},\boldsymbol{\theta}^*) \le 0\right) dP_1(\boldsymbol{z}) \\
&= \Psi(\boldsymbol{\theta}^*).
\end{aligned}
$$

The justification for bringing the limit inside of the integral is due to the dominated convergence theorem. The justification for bringing the limit inside of the indicator function is that the indicator function is continuous everywhere except at 0, which by Assumption 2 occurs with probability 0, and hence does not change the value of the integral. This result establishes that Ψ is continuous in $\boldsymbol{\theta}$.

It is shown in the supplementary materials that $\hat{\theta}_{jn} \xrightarrow{p} \tilde{\theta}_j$, so by Slutsky's theorem it follows that $\hat{\boldsymbol{\theta}}_n \xrightarrow{p} \tilde{\boldsymbol{\theta}}$. Then by the result obtained in equation (A.11), a second application of Slutsky's theorem yields $\Psi(\hat{\boldsymbol{\theta}}_n) \xrightarrow{p} \Psi(\tilde{\boldsymbol{\theta}})$. □

A.2.3 Proof of Theorem 3

Recall the statement of Theorem 3 from the main paper. We will need the following assumptions for the result. Let δ be a small positive constant, and consider the following assumptions.

Assumption 1. F_i^{-1} is a continuous function of $\theta,\ i = 0, 1$.

Assumption 2. $\mathbb{P}\left(\Lambda(Z,\theta) = 0\right) = 0$ for all $\theta \in [\delta, 1-\delta]$.

Theorem 3. *Consider two populations Π_0 and Π_1 with corresponding distribution functions F_0 and F_1 and density functions f_0 and f_1 such that f_0 and f_1 are nonzero on the same domain, and further suppose that Assumptions 1 and 2 hold. Let Z be a random variable with a prior probability π_0 of being a member of population Π_0, and $\pi_1 = 1 - \pi_0$ the prior probability of being a member of population Π_1.*

Next, suppose that there are a finite number of points $x_1 < \cdots < x_t$ such that $F_0(x) = F_1(x)$ for $x \in \{x_1, \ldots, x_t\}$ and that $F_0(x) \neq F_1(x)$ otherwise, and let us denote $x_0 = \inf\{x : f_0(x) > 0\}$ and $x_{t+1} = \sup\{x : f_0(x) > 0\}$. Further assume that for every interval $(x_k,\, x_{k+1})$, that there is exactly one point $z_k^* \in (x_k,\, x_{k+1})$ such that $\pi_0 f_0(z_k^*) = \pi_1 f_1(z_k^*)$, and that $\pi_0 f_0(z) < \pi_1 f_1(z)$ for z on one side of z_k^* and $\pi_0 f_0(z) > \pi_1 f_1(z)$ for z on the other side of z_k^* in the interval.

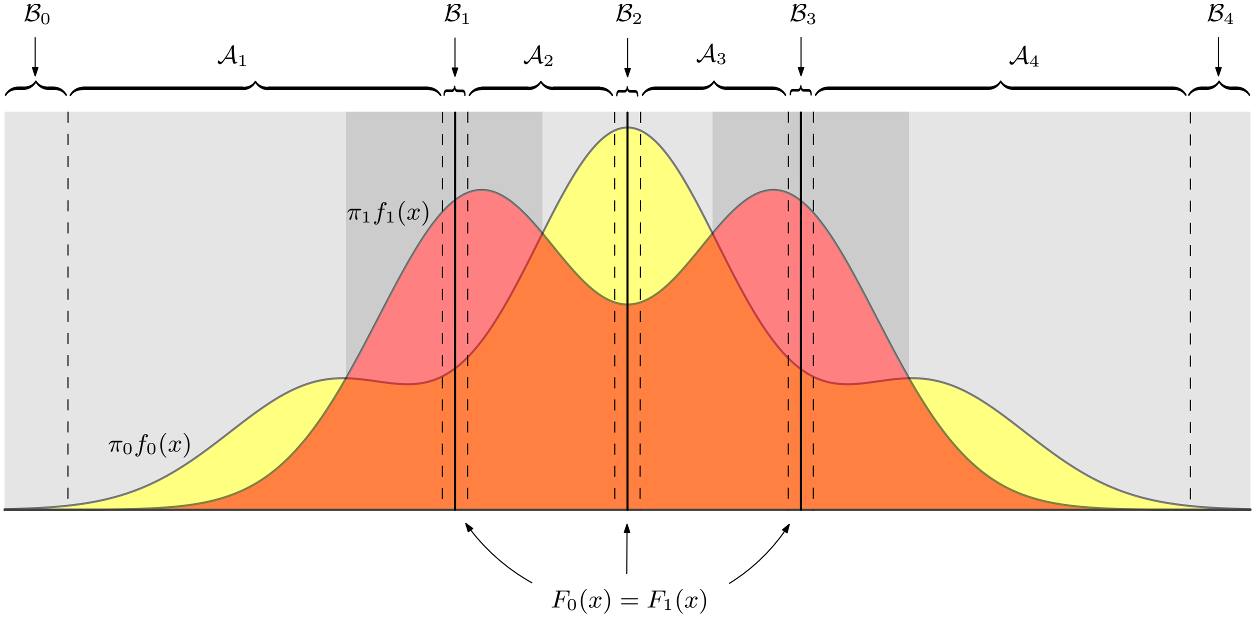
Let $\Psi_n^\dagger$ denote the classification rate of the empirically optimal multimodal quantile classifier, and let $\mathcal{R}$ denote the Bayes rule classification rate. Then under Assumptions 1 and 2, it follows that $\Psi_n^\dagger \xrightarrow{p} \mathcal{R}$.

A.2.3.1 Preliminary propositions

To begin with, we define some terminology and establish some intermediate results that will be of use when proving Theorem 3. We provide the proof of the main theorem in Section A.2.3.2, and then go back and provide the proofs for the intermediate propositions in A.2.3.3.

Let us start by partitioning the domain of Π_0 and Π_1. The idea is to create small intervals around the values of x where $F_0(x) = F_1(x)$ and then have large intervals between those smaller intervals. Consider the setting of Theorem 3 in what follows. Then we have points $x_1 < \cdots < x_t$ such that $F_0(x) = F_1(x)$ for $x \in \{x_1, \ldots, x_t\}$, and points $\tilde{\tau}_1 < \cdots < \tilde{\tau}_{k+1}$ such that $\pi_0 f_0(z) = \pi_1 f_1(z)$ for $z \in \{\tilde{\tau}_1, \ldots, \tilde{\tau}_k\}$, and we additionally define $x_0 = \inf\{x : f_0(x) > 0\}$ and $x_{t+1} = \sup\{x : f_0(x) > 0\}$. Then we consider non-overlapping intervals $\mathcal{B}_0, \ldots, \mathcal{B}_{t+1}$ of the form $(b_k^{low}, b_k^{up}]$ such that $x_k \in \mathcal{B}_k$ and $\tilde{\tau}_k \notin \mathcal{B}_k$, and further define the intervals $\mathcal{A}_1, \ldots, \mathcal{A}_k$ of the form $(a_k^{low}, a_k^{up}]$ such that $a_k^{low} = b_{k-1}^{up}$ and $a_k^{up} = b_k^{low}$. We also at times consider the further subdivision of $\mathcal{A}_k$ into the intervals $\mathcal{A}_k^- = \mathcal{A}_k \cap \{z : z \leq \tilde{\tau}_k\}$ and $\mathcal{A}_k^+ = \mathcal{A}_k \cap \{z : z > \tilde{\tau}_k\}$. See Figure A.3 for an example illustration of such a partitioning.

Figure A.3: An example partitioning of the domain of Π_0 and Π_1 into sets $\mathcal{B}_0$, $\mathcal{A}_1$, $\mathcal{B}_1$, $\mathcal{A}_2$, $\mathcal{B}_2$, $\mathcal{A}_3$, $\mathcal{B}_3$, $\mathcal{A}_4$, and $\mathcal{B}_4$. The partitioning is based on the fact that the $\mathcal{B}_k$ are the regions where it is hard to estimate $\text{sign}(F_1(x) - F_0(x))$.



Next we define a restricted variant of the classification rate for the empirical quantile classifier. Let

$$\Psi^{\dagger}_{n,\mathcal{D}}(\theta) = \pi_0 \int_{\mathcal{D}} \mathbb{1}\left(\Lambda_n(z,\theta) > 0\right) dP_0(z) + \pi_1 \int_{\mathcal{D}} \mathbb{1}\left(\Lambda_n(z,\theta) \leq 0\right) dP_1(z),$$

for $\mathcal{D} \subset \mathbb{R}$. Thus, $\psi^{\dagger}_{n,\mathcal{D}}(\theta)$ is the classification rate for the θ-th empirical quantile classifier with n samples over the region $\mathcal{D}$. We additionally define $\hat{\theta}_{n,\mathcal{D}}$ to be any solution to the equation

$$\hat{\theta}_{n,\mathcal{D}} = \arg\max_{\theta \in T} \Psi_{n,\mathcal{D}}(\theta),$$

where $T = [\delta, 1-\delta]$ for some small positive constant δ, and further define $\hat{\tau}_{n,\mathcal{D}}$ be the decision rule boundary corresponding to the chosen value for $\hat{\theta}_{n,\mathcal{D}}$. We also define a measure of how many more points are correctly classified by the θ-th empirical quantile classifier than are incorrectly classified in the region $\mathcal{D}$ as

$$\begin{aligned}\Gamma_{\mathcal{D}}(\theta) = &\sum_{\substack{z_i \in \Pi_0 \\ z_i \in \mathcal{D}}} \mathbb{1}\left(\Lambda_n(z_i,\theta) > 0\right) + \sum_{\substack{z_i \in \Pi_1 \\ z_i \in \mathcal{D}}} \mathbb{1}\left(\Lambda_n(z_i,\theta) \leq 0\right) \\ &- \sum_{\substack{z_i \in \Pi_0 \\ z_i \in \mathcal{D}}} \mathbb{1}\left(\Lambda_n(z_i,\theta) \leq 0\right) - \sum_{\substack{z_i \in \Pi_1 \\ z_i \in \mathcal{D}}} \mathbb{1}\left(\Lambda_n(z_i,\theta) > 0\right).\end{aligned}$$

With these definitions in hand, we can now begin to present some results. The next proposition deals with estimating the difference between the distribution functions.

Proposition 4. *Suppose that we have sets* $\mathcal{A}_1, \ldots, \mathcal{A}_{t+1}$ *such that* $\inf_{x \in \bigcup \mathcal{A}_k} |F_1(x) - F_0(x)| > 0$. *Then*

$$\lim_{n \to \infty} \mathbb{P}\left(\max_{x \in \bigcup \mathcal{A}_k} \left\{\operatorname{sign}(\hat{F}_{1n_1}(x) - \hat{F}_{0n_0}(x)) - \operatorname{sign}(F_1(x) - F_0(x))\right\} = 0\right) = 1.$$

Next, Supplementary Proposition 5 and Supplementary Corollary 6 provide necessary conditions for the classification rate to differ for the quantile level chosen based on the observations in $\mathcal{A}_k$, compared the quantile level based on the points in the surrounding region selected by the multimodal quantile classifier.

Proposition 5. *Let* $x_1, \ldots, x_{t_n}$ *be the cutpoints chosen by the multimodal empirical quantile classifier, and suppose that* $\operatorname{sign}(\hat{F}_{1n_1}(x) - \hat{F}_{0n_0}(x)) = \operatorname{sign}(F_1(x) - F_0(x))$ *for all* $x \in \mathcal{A}_k$. *Let* k_n *be the index such that* $\mathcal{A}_k \subset (x_{k_n-1}, x_{k_n}]$, *then a necessary condition for the event* $\Psi^{\dagger}_{n,\mathcal{A}_k}(\hat{\theta}_{n,(x_{k_n-1},x_{k_n}]}) \neq \Psi^{\dagger}_{n,\mathcal{A}_k}(\hat{\theta}_{n,\mathcal{A}_k})$ *to occur is that either:*

1. *there is an $a \in \mathbb{R}$ with $x_{k_n-1} < a \leq a_k^{low}$ such that $\Gamma_{(a_k^{low}, \hat{\tau}_{n,\mathcal{A}_k}]}(\hat{\theta}_{n,\mathcal{A}_k}) \leq -\Gamma_{(a, a_k^{low}]}(\hat{\theta}_{n,\mathcal{A}_k})$, or*

2. *there is an $a \in \mathbb{R}$ with $a_k^{up} < a \leq x_{k_n}$ such that $\Gamma_{(\hat{\tau}_{n,\mathcal{A}_k}, a_k^{up}]}(\hat{\theta}_{n,\mathcal{A}_k}) \leq -\Gamma_{[a_k^{up}, a]}(\hat{\theta}_{n,\mathcal{A}_k})$.*

This says that any move of the decision rule boundary from $\mathcal{A}_k$ to one of the surrounding intervals enclosed by $(x_{k_n-1}, x_{k_n}]$ requires an interval directly bordering $\mathcal{A}_k$ so that changing the boundary would cause enough of a gain in the number of correctly classified points in the bordering interval to outweigh the decrease in correctly classified points in $\mathcal{A}_k$. The following corollary provides a similar result, but now with respect to the true optimal decision rule boundary instead of the estimated boundary.

Corollary 6 (Corollary to Supplementary Proposition 5)**.** *Consider the setting of Supplementary Proposition 5, and let $\tilde{\theta}_k$ and $\tilde{\tau}_k$ be the optimal quantile level and corresponding decision rule boundary for the subproblem restricted to $\mathcal{A}_k$. Then a necessary condition for the event $\Psi^{\dagger}_{n,\mathcal{A}_k}(\hat{\theta}_{n,(x_{k_n-1}, x_{k_n}]}) \neq \Psi^{\dagger}_{n,\mathcal{A}_k}(\tilde{\theta}_k)$ to occur is that either:*

1. *there is an $a \in \mathbb{R}$ with $x_{k_n-1} < a \leq a_k^{low}$ such that $\Gamma_{(a_k^{low}, \tilde{\tau}_k]}(\tilde{\theta}_k) \leq -\Gamma_{(a, a_k^{low}]}(\tilde{\theta}_k)$, or*

2. *there is an $a \in \mathbb{R}$ with $a_k^{up} < a \leq x_{k_n}$ such that $\Gamma_{[\tilde{\tau}_k, a_k^{up}]}(\tilde{\theta}_k) \leq -\Gamma_{[a_k^{up}, a]}(\tilde{\theta}_k)$.*

Next we establish a general convergence result that will be of use later.

Proposition 7. *Let $G_1, G_2, \ldots$, and $H_1, H_2, \ldots$ be two sequences of events such that G_k and H_k corresponding to the same probability space for each k. Suppose that $\lim_{n\to\infty} \mathbb{P}(G_n) = \pi$, and $\lim_{n\to\infty} \mathbb{P}(H_n) = 1$. Then $\lim_{n\to\infty} \mathbb{P}(G_n | H_n) = \pi$.*

Finally, the following result says that we can construct the partitioning of the populations' domains so that conditional on the estimated distribution functions having the same within-region stochastic ordering for $\mathcal{A}_1, \ldots, \mathcal{A}_{t+1}$, then the classification rate for the quantile classifier based on $\mathcal{A}_k$ is equal to the classification rate for the quantile classifier based on the surrounding subproblem chosen by the classifier with high probability.

Proposition 8. *Consider the setting of Theorem 3, and let H_n denote the event that $\operatorname{sign}(\hat{F}_{1n_1}(x) - \hat{F}_{0n_0}(x)) = \operatorname{sign}(F_1(x) - F_0(x))$ for all $x \in \mathcal{A}_k$. Then we can choose $\mathcal{B}_0, \mathcal{A}_1, \mathcal{B}_2, \ldots, \mathcal{A}_t, \mathcal{B}_{t+1}$ so that $\lim_{n\to\infty} \mathbb{P}\left(\Psi^{\dagger}_{n,\mathcal{A}_k}(\hat{\theta}_{n,(x_{k_n-1}, x_{k_n}]}) = \Psi^{\dagger}_{n,\mathcal{A}_k}(\hat{\theta}_{n,\mathcal{A}_k}) \,\middle|\, H_n\right) = 1$ for $k = 1, \ldots, t+1$.*

A.2.3.2 Proof of Theorem 3

Proof of Theorem 3. Fix $\epsilon, \delta > 0$. Then we want to show that there exists some $N_0 \in \mathbb{N}$ such that for all $N \geq N_0$, then

$$\mathbb{P}\left(\left|\Psi_N^\dagger - \mathcal{R}\right| > \epsilon\right) < \delta.$$

In what follows we let $\Psi_{N,\mathcal{D}}^\dagger$ denote the classification rate for a subset of the domain $\mathcal{D}$, but using the classification rule constructed by the multimodal quantile classifier for the entire domain. Similarly, we let $\mathcal{R}_\mathcal{D}$ denote the Bayes rule classification rate restricted to the domain $\mathcal{D}$. We will assume that there are $t-1$ values of x such that $F_0(x) = F_1(x)$, and let us choose $\mathcal{B}_0, \dots, \mathcal{B}_t$ so that $\sum_{\ell=0}^t \mathcal{R}_{\mathcal{B}_\ell} \leq \frac{\epsilon}{2}$. Then we have

$$\begin{aligned}
&\mathbb{P}\left(\left|\Psi_N^\dagger - \mathcal{R}\right| > \epsilon\right) \\
&\quad = \mathbb{P}\left(\sum_{k=1}^t \left|\Psi_{N,\mathcal{A}_k}^\dagger - \mathcal{R}_{\mathcal{A}_k}\right| + \sum_{\ell=0}^t \left|\Psi_{N,\mathcal{B}_\ell}^\dagger - \mathcal{R}_{\mathcal{B}_\ell}\right| > \epsilon\right) \\
&\quad \leq \mathbb{P}\left(\left\{\sum_{k=1}^t \left|\Psi_{N,\mathcal{A}_k}^\dagger - \mathcal{R}_{\mathcal{A}_k}\right| > \frac{\epsilon}{2}\right\} \cap \left\{\sum_{\ell=0}^t \left|\Psi_{N,\mathcal{B}_\ell}^\dagger - \mathcal{R}_{\mathcal{B}_\ell}\right| > \frac{\epsilon}{2}\right\}\right) \\
&\quad \leq \mathbb{P}\left(\sum_{k=1}^t \left|\Psi_{N,\mathcal{A}_k}^\dagger - \mathcal{R}_{\mathcal{A}_k}\right| > \frac{\epsilon}{2}\right) + \mathbb{P}\left(\sum_{\ell=0}^t \left|\Psi_{N,\mathcal{B}_\ell}^\dagger - \mathcal{R}_{\mathcal{B}_\ell}\right| > \frac{\epsilon}{2}\right) \\
&\quad = \mathbb{P}\left(\sum_{k=1}^t \left|\Psi_{N,\mathcal{A}_k}^\dagger - \mathcal{R}_{\mathcal{A}_k}\right| > \frac{\epsilon}{2}\right) \qquad \text{(A.12)} \\
&\quad \leq \mathbb{P}\left(\bigcup_{k=1}^t \left\{\left|\Psi_{N,\mathcal{A}_k}^\dagger - \mathcal{R}_{\mathcal{A}_k}\right| > \frac{\epsilon}{2t}\right\}\right) \\
&\quad \leq \sum_{k=1}^t \mathbb{P}\left(\left|\Psi_{N,\mathcal{A}_k}^\dagger - \mathcal{R}_{\mathcal{A}_k}\right| > \frac{\epsilon}{2t}\right) \\
&\quad = \sum_{k=1}^t \left\{1 - \mathbb{P}\left(\left|\Psi_{N,\mathcal{A}_k}^\dagger - \mathcal{R}_{\mathcal{A}_k}\right| \leq \frac{\epsilon}{2t}\right)\right\}.
\end{aligned}$$

We note that the equality for (A.12) is due to the construction of the $\mathcal{B}_\ell$. Thus, it suffices to show that

$$\mathbb{P}\left(\left|\Psi_{N,\mathcal{A}_k} - \mathcal{R}_{\mathcal{A}_k}\right| \leq \frac{\epsilon}{2t}\right) > 1 - \delta_1,$$

for $k = 1, \dots, t$, and where $\delta_1 = \frac{\delta}{t}$.

Next, we define a variant of $\Psi^{\dagger}_{N,\mathcal{D}}$ that we call $\Psi^{\dagger *}_{N,\mathcal{D}}$, and which is defined as the classification rate for the regular (non-multimodal) quantile classifier where the empirically optimal quantile level is chosen based only on the observations in $\mathcal{D}$. We also define H_N to be the event that $\operatorname{sign}(\hat{F}_{1n}(x) - \hat{F}_{0n}(x)) = \operatorname{sign}(F_1(x) - F_0(x))$ for all $x \in \mathcal{A}_k$ and for $k = 1, \dots, t$. Then for large N we have

$$\begin{aligned}
&\mathbb{P}\left(\left|\Psi^{\dagger}_{N,\mathcal{A}_k} - \mathcal{R}_{\mathcal{A}_k}\right| \le \frac{\epsilon}{2t}\right) \\
&\quad \ge \mathbb{P}\left(\left\{\left|\Psi^{\dagger *}_{N,\mathcal{A}_k} - \mathcal{R}_{\mathcal{A}_k}\right| \le \frac{\epsilon}{2t}\right\} \cap \left\{\Psi^{\dagger *}_{N,\mathcal{A}_k} = \Psi^{\dagger}_{N,\mathcal{A}_k}\right\} \cap H_N\right) \qquad \text{(A.13)} \\
&\quad = 1 - \mathbb{P}\left(\left\{\left|\Psi^{\dagger *}_{N,\mathcal{A}_k} - \mathcal{R}_{\mathcal{A}_k}\right| \le \frac{\epsilon}{2t}\right\}^c \cup \left\{\left\{\Psi^{\dagger *}_{N,\mathcal{A}_k} = \Psi^{\dagger}_{N,\mathcal{A}_k}\right\} \cap H_N\right\}^c\right) \\
&\quad = 1 - \mathbb{P}\left(\left\{\left|\Psi^{\dagger *}_{N,\mathcal{A}_k} - \mathcal{R}_{\mathcal{A}_k}\right| \le \frac{\epsilon}{2t}\right\}^c\right) - \mathbb{P}\left(\left\{\left\{\Psi^{\dagger *}_{N,\mathcal{A}_k} = \Psi^{\dagger}_{N,\mathcal{A}_k}\right\} \cap H_N\right\}^c\right) \\
&\quad > 1 - \frac{\delta_1}{2} - \mathbb{P}\left(\left\{\left\{\Psi^{\dagger *}_{N,\mathcal{A}_k} = \Psi^{\dagger}_{N,\mathcal{A}_k}\right\} \cap H_N\right\}^c\right),
\end{aligned}$$

where the inequality for (A.13) is due to the convergence result in Hennig and Viroli (2016). Furthermore,

$$\begin{aligned}
&\mathbb{P}\left(\left\{\left\{\Psi^{\dagger *}_{N,\mathcal{A}_k} = \Psi^{\dagger}_{N,\mathcal{A}_k}\right\} \cap H_N\right\}^c\right) \\
&\quad = 1 - \mathbb{P}\left(\left\{\Psi^{\dagger *}_{N,\mathcal{A}_k} = \Psi^{\dagger}_{N,\mathcal{A}_k}\right\} \cap H_N\right) \\
&\quad = 1 - \mathbb{P}\left(\Psi^{\dagger *}_{N,\mathcal{A}_k} = \Psi^{\dagger}_{N,\mathcal{A}_k} \mid H_N\right) \mathbb{P}\left(H_N\right),
\end{aligned}$$

so it suffices to show that

$$\mathbb{P}\left(\Psi^{\dagger *}_{N,\mathcal{A}_k} = \Psi^{\dagger}_{N,\mathcal{A}_k} \mid H_N\right) > 1 - \delta_2 \qquad \text{and} \qquad \mathbb{P}\left(H_N\right) > 1 - \delta_2, \qquad \text{(A.14)}$$

and where $1 - \delta_2 = \left(1 - \frac{\delta_1}{2}\right)^{1/2}$. But we know that both inequalities in (A.14) hold for large N due to Supplementary Proposition 8 and Proposition 4, so this establishes the main result of the theorem. □

A.2.3.3 Proofs of preliminary propositions

Proof of Supplementary Proposition 4. From the Glivenko-Cantelli theorem (e.g. see Van der Vaart (2000)), we know that for every $\epsilon, \delta > 0$ there exists an $N_0 \in \mathbb{N}$ such that for every $N \geq N_0$

$$\mathbb{P}\left(\sup_x \left|\hat{F}_{0N}(x) - F_0(x)\right| > \frac{\epsilon}{2}\right) < \frac{\delta}{2},$$

and additionally there exists an $N_1 \in \mathbb{N}$ such that for every $N \geq N_1$

$$\mathbb{P}\left(\sup_x \left|\hat{F}_{1N}(x) - F_1(x)\right| > \frac{\epsilon}{2}\right) < \frac{\delta}{2}.$$

Then for $N \geq \max(N_0, N_1)$, we have

$$\begin{aligned}
&\mathbb{P}\left(\sup_x \left|\hat{F}_{1N}(x) - \hat{F}_{0N}(x) - (F_1(x) - F_0(x))\right| > \epsilon\right) \\
&\quad = \mathbb{P}\left(\sup_x \left|\hat{F}_{1N}(x) - F_1(x) + F_0(x) - \hat{F}_{0N}(x)\right| > \epsilon\right) \\
&\quad \leq \mathbb{P}\left(\sup_x \left\{\left|\hat{F}_{1N}(x) - F_1(x)\right| + \left|\hat{F}_{0N}(x) - F_0(x)\right|\right\} > \epsilon\right) \\
&\quad \leq \mathbb{P}\left(\sup_x \left|\hat{F}_{1N}(x) - F_1(x)\right| + \sup_x \left|\hat{F}_{0N}(x) - F_0(x)\right| > \epsilon\right) \\
&\quad \leq \mathbb{P}\left(\left\{\sup_x \left|\hat{F}_{1N}(x) - F_1(x)\right| > \frac{\epsilon}{2}\right\} \cup \left\{\sup_x \left|\hat{F}_{0N}(x) - F_0(x)\right| > \frac{\epsilon}{2}\right\}\right) \\
&\quad \leq \mathbb{P}\left(\sup_x \left|\hat{F}_{1N}(x) - F_1(x)\right| > \frac{\epsilon}{2}\right) + \mathbb{P}\left(\sup_x \left|\hat{F}_{0N}(x) - F_0(x)\right| > \frac{\epsilon}{2}\right) \\
&\quad \leq \frac{\delta}{2} + \frac{\delta}{2} \\
&\quad = \delta,
\end{aligned}$$

Suppose now that $\inf_{x \in \bigcup \mathcal{A}_k} |F_1(x) - F_0(x)| = \kappa$ for $\kappa > 0$. Then

$$\begin{aligned}
&\mathbb{P}\left(\max_{x \in \bigcup \mathcal{A}_k} \left\{\operatorname{sign}(\hat{F}_{1n}(x) - \hat{F}_{0n}(x)) - \operatorname{sign}(F_1(x) - F_0(x))\right\} = 0\right) \\
&\quad \leq \mathbb{P}\left(\sup_x \left|\hat{F}_{1N}(x) - \hat{F}_{0N}(x) - (F_1(x) - F_0(x))\right| \geq \kappa\right). \qquad \text{(A.15)}
\end{aligned}$$

But since for every $\delta > 0$ we can find an $N^* \in \mathbb{N}$ such that for every $N \geq N^*$ the expression in (A.15) is less than κ the definition of a limit is satisfied and the result is shown. □

Proof of Supplementary Proposition 5. Firstly, we note that if $\Psi^{\dagger}_{n,\mathcal{A}_k}(\hat{\theta}_{n,(x_{k_n-1},x_{k_n}]}) \neq \Psi^{\dagger}_{n,\mathcal{A}_k}(\hat{\theta}_{n,\mathcal{A}_k})$ then $\hat{\tau}_{n,(x_{k_n-1},x_{k_n}]} \notin \mathcal{A}_k$. For if this were indeed the case, then we would have a suboptimal classification of points inside of $\mathcal{A}_k$ without changing the classification of points in the surrounding intervals.

Suppose then that $\hat{\tau}_{n,(x_{k_n-1},x_{k_n}]} \in (x_{k_n-1}, a_k^{low}]$; let us call this value and the corresponding quantile level τ^* and θ^* for ease of notation. This choice of quantile level reverses the classification of points in the region $(a_k^{low}, \tau^*]$ while leaving the classifications the same in $(\tau^*, x_{k_n}]$, so that there are now $\Gamma_{(a_k^{low},\tau^*]}(\theta^*)$ fewer correct classifications in those regions. Thus, we need to get at least that many more correct in $(\tau^*, a_k]$ to counteract this loss, since the classification of observations in $(x_{k_n-1}, \tau^*]$ remains unchanged.

A similar argument can be made for $\hat{\tau}_{n,(x_{k_n-1},x_{k_n}]} \in (a_k^{up}, x_{k_n}]$. These statements in conjunction verify Supplementary Proposition 5. □

Proof of Corollary 6. We note that $\Gamma_{(a_k^{low},\tilde{\tau}_k)}(\tilde{\theta}_k) \leq \Gamma_{(a_k^{low},\hat{\tau}_{n,\mathcal{A}_k}]}(\hat{\theta}_{n,\mathcal{A}_k})$ and that $\Gamma_{[\tilde{\tau}_k,a_k^{up}]}(\tilde{\theta}_k) \leq \Gamma_{(\hat{\tau}_{n,\mathcal{A}_k},a_k^{up}]}$ $(\hat{\theta}_{n,\mathcal{A}_k})$, from which it follows that both conditions in Corollary 6 are weaker then the conditions in Supplementary Proposition 5. Since they are necessary conditions in the proposition, then they must also be necessary conditions in the corollary. □

Proof of Supplementary Proposition 7. We have

$$\begin{aligned}
&\lim_{n\to\infty} \mathbb{P}\left(A_n | H_n\right) \\
&\quad = \lim_{n\to\infty} \frac{\mathbb{P}\left(A_n \cap H_n\right)}{\mathbb{P}\left(H_n\right)} \\
&\quad = \frac{\lim_{n\to\infty} \mathbb{P}\left(A_n \cap H_n\right)}{\lim_{n\to\infty} \mathbb{P}\left(H_n\right)} \\
&\quad = \lim_{n\to\infty} \mathbb{P}\left(A_n \cap H_n\right) \\
&\quad = \lim_{n\to\infty} \left\{1 - \mathbb{P}\left(A_n^c \cup H_n^c\right)\right\} \\
&\quad \geq \lim_{n\to\infty} \left\{1 - \left[\mathbb{P}\left(A_n^c\right) + \mathbb{P}\left(H_n^c\right)\right]\right\} \\
&\quad = \lim_{n\to\infty} \left\{1 - \left[1 - \mathbb{P}\left(A_n\right) + 1 - \mathbb{P}\left(H_n\right)\right]\right\}
\end{aligned}$$

$$\begin{aligned}
&= \lim_{n\to\infty} \Big\{ \mathbb{P}\left(A_n\right) - 1 + \mathbb{P}\left(H_n\right) \Big\} \\
&= \lim_{n\to\infty} \mathbb{P}\left(A_n\right) - 1 + \mathbb{P}\left(H_n\right) \\
&= \pi - 1 + 1 \\
&= \pi.
\end{aligned}$$

Furthermore,

$$\begin{aligned}
&\lim_{n\to\infty} \mathbb{P}\left(A_n | H_n\right) \\
&\qquad = \lim_{n\to\infty} \frac{\mathbb{P}\left(A_n \cap H_n\right)}{\mathbb{P}\left(H_n\right)} \\
&\qquad = \frac{\lim_{n\to\infty} \mathbb{P}\left(A_n \cap H_n\right)}{\lim_{n\to\infty} \mathbb{P}\left(H_n\right)} \\
&\qquad = \lim_{n\to\infty} \mathbb{P}\left(A_n \cap H_n\right) \\
&\qquad \leq \lim_{n\to\infty} \mathbb{P}\left(A_n\right) \\
&\qquad = \pi.
\end{aligned}$$

Thus, by putting these two pieces together, we obtain the desired result. □

Proof of Supplementary Proposition 8. Let us focus our attention for the time-being to the region $\mathcal{A}_k$ and its surrounding regions $\mathcal{B}_{k-1}$ and $\mathcal{B}_k$. Then a necessary condition for the event $\Psi^{\dagger}_{n,\mathcal{A}_k}(\hat{\theta}_{n,(x_{k_n-1},x_{k_n}]})$ $\neq \Psi^{\dagger}_{n,\mathcal{A}_k}(\hat{\theta}_{n,\mathcal{A}_k})$ to occur is given by Corollary 6. We will show in what follows that $\mathcal{B}_{k-1}, \mathcal{A}_k$, and $\mathcal{B}_k$ can be chosen so that the total number of observations in $\mathcal{B}_{k-1}$ is less than $\Gamma_{(a_k,\tilde{\tau}_k)}(\tilde{\theta}_k)$ with high probability, and the total number of observations in $\mathcal{B}_k$ is less than $\Gamma_{(\tilde{\tau}_k,b_k)}(\tilde{\theta}_k)$ with high probability, which together implies that the necessary condition in Corollary 6 does not occur.

Let us define the random variables $U^{(i)}_\ell = \mathbb{1}\left(X_\ell \text{ drawn from } \Pi_i, X_\ell \in \mathcal{A}^+_k\right),\ i = 0, 1$ and $V_\ell = \mathbb{1}\left(X_\ell \in \mathcal{B}_k\right)$. Further define the quantities $\pi_{U^{(i)}} = \mathbb{E}\, U^{(i)}_\ell,\ i = 0, 1$ and $\pi_V = \mathbb{E}\, V_\ell$. Now, under the assumptions of Theorem 3 we have either $\pi_0 f_0(x) > \pi_1 f_1(x)$ or $\pi_0 f_0(x) < \pi_1 f_1(x)$ for all $x \in \mathcal{A}^+_k$. Suppose for the time-being that it is the latter, then a result of $\sum_{\ell=1}^N \left(U^{(1)}_\ell - U^{(0)}_\ell - V_\ell\right) > 0$ means that part 2 of the necessary condition in Corollary 6 cannot occur since the value of $\sum_{\ell=1}^{N} V_\ell \geq -\Gamma_{[a_k^{up},a]}(\tilde{\theta}_k)$ for any $a \in \mathbb{R}$ with $a_k^{up} < a \leq x_{k_n}$.

Now, since we are considering the case where $\pi_0 f_0(x) < \pi_1 f_1(x)$ for all $x > \tilde{\tau}_k$, it follows that $\pi_{U^{(1)}} > \pi_{U^{(0)}}$ for any choice of $\mathcal{A}_k^+$. This being the case, let us choose $\mathcal{A}_k^+$ and $\mathcal{B}_k$ such that $\pi_V < \pi_{U^{(1)}} - \pi_{U^{(0)}}$. Then we have

$$\mathbb{E}\left[\sum_{\ell=1}^{N}\Big(U_\ell^{(1)} - U_\ell^{(0)} - V_\ell\Big)\right] = n\Big(\pi_{U^{(0)}} - \pi_{U^{(1)}} - \pi_V\Big) > 0. \tag{A.16}$$

For the case when $\pi_0 f_0(x) > \pi_1 f_1(x)$ then we would instead consider $\sum_{\ell=1}^{N}\big(U_\ell^{(0)} - U_\ell^{(1)} - V_\ell\big)$. Next, since $U_\ell^{(1)}, U_\ell^{(0)}$, and V_ℓ along with other possible values for X_ℓ form a portion of a multivariate distribution, we have

$$\begin{aligned}
&\text{Var}\left[\sum_{\ell=1}^{N}\Big(U_\ell^{(1)} - U_\ell^{(0)} - V_\ell\Big)\right] \qquad \text{(A.17)}\\
&\quad = n\,\text{Var}\left[U_1^{(1)} - U_1^{(0)} - V_1\right]\\
&\quad = n\Big(\text{Var}\left[U_1^{(1)}\right] + \text{Var}\left[U_1^{(0)}\right] + \text{Var}\left[V_1\right]\\
&\qquad\quad - \text{Cov}\left[U_1^{(1)}, U_1^{(0)}\right] - \text{Cov}\left[U_1^{(1)}, V_1\right] + \text{Cov}\left[U_1^{(0)}, V_1\right]\Big)\\
&\quad = n\Big(\pi_{U^{(1)}}(1-\pi_{U^{(1)}}) + \pi_{U^{(0)}}(1-\pi_{U^{(0)}}) + \pi_V(1-\pi_V)\\
&\qquad\quad + \pi_{U^{(1)}}\pi_{U^{(0)}} + \pi_{U^{(1)}}\pi_V - \pi_{U^{(0)}}\pi_V\Big)\\
&\quad < \infty.
\end{aligned}$$

Let us denote $W_\ell = U_\ell^{(1)} - U_\ell^{(0)} - V_\ell$, and let $\mu_W = \mathbb{E}\,W$ and $\sigma_W^2 = \text{Var}\,[W]$, then

$$\begin{aligned}
&\mathbb{P}\left(\sum_{\ell=1}^{N} W_\ell \leq 0\right)\\
&\quad = \mathbb{P}\left(\sum_{\ell=1}^{N} W_\ell - n\mu_W \leq -n\mu_W\right)\\
&\quad = \mathbb{P}\left(-\left(\sum_{\ell=1}^{N} W_\ell - n\mu_W\right) \geq n\mu_W\right)
\end{aligned}$$

$$
\begin{aligned}
&\leq \mathbb{P}\left(\left|\sum_{\ell=1}^{N} W_\ell - n\mu_W\right| \geq n\mu_W\right)\\
&= \mathbb{P}\left(\frac{\left|\sum_{\ell=1}^{N} W_\ell - n\mu_W\right|}{\sqrt{n}\sigma_W} \geq \frac{n\mu_W}{\sqrt{n}\sigma_W}\right)\\
&= \mathbb{P}\left(\frac{\left|\sum_{\ell=1}^{N} W_\ell - n\mu_W\right|}{\sqrt{n}\sigma_W} \geq \frac{\sqrt{n}\mu_W}{\sigma_W}\right)\\
&\leq \frac{\sigma_W^2}{n\mu_W^2}, \qquad \text{(A.18)}
\end{aligned}
$$

where the final inequality is due to Chebyshev's inequality. We note that this upper bound can also be obtained for an analogous statement regarding the number of points in $\mathcal{B}_{k-1}$ compared to the number of correctly classified points minus the number of incorrectly classified points in $\mathcal{A}_k^-$.

Let us denote the event described in (A.17) and the analogous event for $\mathcal{A}_k^-$ and $\mathcal{B}_{k-1}$ as $S_{n,k}^+$ and $S_{n,k}^-$, respectively. Then we have

$$
\begin{aligned}
&\mathbb{P}\left(\Psi^\dagger_{n,\mathcal{A}_k}(\hat{\theta}_{n,(x_{k_n-1},x_{k_n}]}) \neq \Psi^\dagger_{n,\mathcal{A}_k}(\hat{\theta}_{n,\mathcal{A}_k}) \bigm| H_N\right)\\
&\qquad \leq \mathbb{P}\left(S_{n,k}^- \cup S_{n,k}^+ \bigm| H_N\right)\\
&\qquad \leq \mathbb{P}\left(S_{n,k}^- \bigm| H_N\right) + \mathbb{P}\left(S_{n,k}^+ \bigm| H_N\right). \qquad \text{(A.19)}
\end{aligned}
$$

Now from (A.18) and the analogous result for $\mathcal{A}_k^-$ and $\mathcal{B}_{k-1}$, we obtain that $\lim_{n\to\infty} \mathbb{P}\left(S_{n,k}^-\right) = \lim_{n\to\infty} \mathbb{P}\left(S_{n,k}^+\right) = 0$. Furthermore, from Supplementary Proposition 7, we obtain that $\lim_{n\to\infty} \mathbb{P}\left(S_{n,k}^- \bigm| H_N\right) = \lim_{n\to\infty} \mathbb{P}\left(S_{n,k}^+ \bigm| H_N\right) = 0$, which in turn forces (A.19) to 0 in the limit.

Thus we conclude that $\lim_{n\to\infty} \mathbb{P}\left(\Psi^\dagger_{n,\mathcal{A}_k}(\hat{\theta}_{n,(x_{k_n-1},x_{k_n}]}) \neq \Psi^\dagger_{n,\mathcal{A}_k}(\hat{\theta}_{n,\mathcal{A}_k}) \bigm| H_N\right)$ for appropriately chosen $\mathcal{B}_{k-1}, \mathcal{A}_k$, and $\mathcal{B}_k$. Furthermore, we can choose $\mathcal{B}_0, \mathcal{A}_1, \mathcal{B}_1, \ldots, \mathcal{A}_{t+1}, \mathcal{B}_{t+1}$ so that this expression holds for all k, since it only requires making the $\mathcal{B}_k$ small enough so that the expression in (A.16) and it's counterpart hold for each k. Since we can make the $\mathcal{B}_k$ arbitrarily small, we can ensure that it is small enough to satisfy such an expression for both (or just one in the case of $\mathcal{B}_0$ or $\mathcal{B}_{t+1}$) of the neighboring $\mathcal{A}_k$. □

A.3 Extended simulation results

A.3.1 Simulated data scenarios

Consider the following framework for each of the scenarios. Let $\boldsymbol{Z} = (Z_1, \ldots, Z_p) \sim N(\mathbf{0}, \boldsymbol{\Sigma})$, and let $\boldsymbol{V}_i = (V_{i1}, \ldots, V_{ip})$ and $\boldsymbol{W}_i = (W_{i1}, \ldots, W_{ip})$ be observations from two populations, $i = 1, \ldots, n/2$, and where the $\boldsymbol{V}_i$ and $\boldsymbol{W}_i$ are all mutually independent. There are two choices of covariance matrices considered for $\boldsymbol{\Sigma}$. In some scenarios we consider independent data where the covariance matrix is the identity matrix, and in other scenarios we consider the correlated data with an autoregressive lag 1 covariance matrix (abbreviated as AR1). The AR1 matrix is specified with variance parameters identically 1 on the diagonal, and correlation parameter 0.8. In the interest of continuity, the scenarios are designed to be similar to the simulation settings presented in Hennig and Viroli (2016).

In the first and second scenarios, we consider two classes each with features drawn from a multivariate Gaussian distribution and with one class given a location shift. That is to say that for each feature we consider $V_{ij} \sim Z_j$ and $W_{ij} \sim Z_j + 0.35$. In the first scenario we consider uncorrelated data for the underlying Gaussian distribution, and in the second scenario we consider autoregressive correlation.

In the third and fourth scenarios, we consider highly skewed data by exponentiating the components of a Gaussian distribution. That is to say that for each feature we consider $V_{ij} \sim \exp(Z_j)$ and $W_{ij} \sim \exp(Z_j) + 0.35$. In the third scenario we consider uncorrelated data for the underlying Gaussian distribution, and in the fourth scenario we consider autoregressive correlation.

In the fifth and sixth scenarios we consider different distributions for each of 5 blocks of features. In more detail, the data is sampled from a Gaussian distribution and then p features are evenly split into 5 blocks, with the following transformations performed to the various blocks: (i) $V_{ij} \sim Z_j$ and $W_{ij} \sim Z_j + 0.2$, (ii) $V_{ij} \sim \exp(Z_j)$ and $W_{ij} \sim \exp(Z_j) + 0.2$, (iii) $V_{ij} \sim \log|Z_j|$ and $W_{ij} \sim \log|Z_j| + 0.1$, (iv) $V_{ij} \sim Z_j^2$ and $W_{ij} \sim Z_j^2 + 0.2$, and (v) $V_{ij} \sim |Z_j|^{1/2}$ and $W_{ij} \sim |Z_j|^{1/2} + 0.1$. In the fifth scenario we consider uncorrelated data for the underlying Gaussian distribution, and in the sixth scenario we consider autoregressive correlation.

In the seventh and eight scenarios, we considered data with different distributional shapes within each feature. In the seventh scenario we considered data with independent beta distributed features and in the eight scenario we considered data with independent gamma distributed features. The beta distribution parameters were sampled as follows. Two shape parameters were sampled for each feature each from a $\mathit{unif}(0.1, 3)$ and $\mathit{unif}(0.5, 3)$ distribution for each shape parameter, respectively. Then for each class and feature, each parameter was transformed by taking the absolute value of some additive Gaussian random noise. So for example, suppose that α_j and β_j are the shape parameters drawn for the j-th feature. Then for each class the distributional parameters are each sampled from $|\alpha_j + N(0, \sigma_j^2)|$ and $|\beta_j + N(0, \sigma_j^2)|$ distributions. The parameters $\sigma_1, \ldots, \sigma_{50}$ were given by a fixed sequence each with values between 0.05 and 0.20. Once the shape parameters were sampled, the same parameters were used for every replicate and every simulation study. Finally, the gamma distribution parameters in the eighth scenario were sampled in essentially the same manner.

In the ninth scenario we consider a setting suggested to us by one the reviewers in which feature deletion occurs for marginal-based methods such as composite quantile-based classifiers. We consider two classes that are drawn from a multivariate Gaussian distribution, one with a mean of $\mathbf{0}$ and one with a mean vector $\boldsymbol{\mu}$ such that $\mu_1 = \mu_3 = \cdots = 0.5$ and $\mu_2 = \mu_4 = \cdots = 0$. Both classes' distributions are equipped with the same AR1 covariance matrix $\mathbf{\Sigma}$ with variance parameters identically 1 on the diagonal and a correlation parameter of 0.8.

In the tenth scenario, we consider data with features that follow independent mixture Gaussian distributions. The first population is a mixture of three Gaussian distribution each with variance 1 and with means -3, 0, and 3, and prior probabilities of 0.2, 0.6, and 0.2 respectively. The second population is a mixture of two Gaussian distributions each with variance 1 and with means -1.5 and 1.5, and prior probabilities of 0.5, and 0.5 respectively. This is the setting shown in Table A.8.

A.3.2 Classifier implementations and settings

We compared the misclassification rate from the composite quantile-based classifiers model with that of nine other classification methods: quantile-based classifiers Hennig and Viroli (2016), FANS Fan et al. (2016), penalized linear regression (ℓ_1 penalty) Park and Hastie (2007), support vector machine (radial kernel) Cortes and Vapnik (1995), k-nearest neighbor Cover and Hart (1967), naive Bayes Hastie et al. (2009), nearest shrunken centroids Tibshirani et al. (2002), penalized LDA Witten and Tibshirani (2011), and decision trees Breiman et al. (1984).

The composite quantiles-based classifier is implemented using the R programming language R Core Team (2016); the source code is available from the authors upon request. Quantile-based methods is implemented as the R package `quantileDA`. There are several methods of quantifying distributional skew provided by the package; we used the default Galton skewness measure. FANS is implemented as MATLAB MATLAB

(2016) source code and is available upon request by the authors of Fan et al. (2016). Results from both the FANS method and FANS2 method are shown since FANS2 is similar to the augmented version of composite quantile-based classifiers. Penalized linear regression is implemented as the R package `glmnet` and uses the ℓ_1 penalty with 10-fold cross validation to select the penalty parameter. Support vector machine is implemented in the R package `e1071` through the C++ library `libsvm`. The Gaussian kernel was used, with tuning parameters for each simulation selected using the function `tune.svm` over the kernel coefficient parameter from among $\{0.001, 0.01, 0.1, 1, 2\}$, and the constraints violation cost from among $\{1, 2, 4, 8, 16\}$. k-nearest neighbors is implemented in the R package `class` and uses leave-one-out cross validation to choose the number of neighbors considered from among $\{1, \ldots, 9\}$. Naive Bayes is implemented in the R package `e1071`. Nearest shrunken centroids is implemented as the R package `pamr` with 10-fold cross validation used to select the threshold parameter. The version of penalized linear discriminant analysis compared in the simulation studies is that proposed in Witten and Tibshirani (2011), and is implemented as the R package `PenalizedLDA` using 6-fold cross validation. Decision trees is implemented as the R package `rpart`. All packages used in the numerical analysis are available from the Comprehensive R Archive Network.

A.3.3 Simulation results

The composite quantile-based classifier is abbreviated as CQC. Results for the second variant of composite quantile-based classifiers where the transformed quantile distances data is augmented by the original data are also shown and are abbreviated as CQC augmented.

Simulation results for the Gaussian setting are shown in Table A.1. We would expect composite quantile-based classifiers to be suboptimal in this setting compared to LDA and similar classifiers since they do not make use of the distributional information, and are interested in the magnitude of the loss of efficiency. One thing to notice is that the augmented version of CQC performs better nearly everywhere as compared to the non-augmented version. This is to be expected since the Bayes rule decision boundary is linear in the original features, and is also the case for FANS2 as compared to FANS. In this setting, nearest shrunken centroids and penalized LDA show the best performance. When n is large compared to p, marginal quantile-based classifiers perform reasonably well, although performance deteriorates relative to other methods when conditions are less ideal.

Simulation results for the exponentiated Gaussian setting are shown in Table A.2. In this setting the 0-th quantile level for all features is optimal and quantile-based perform extremely well. CQC performs well also, although never as well as quantile-based classifiers. We see this as being due to several reasons. Firstly, quantile-based classifiers composite quantile-based classifiers sacrifice part of the data set in order to train the linear combination coefficients; this has the effect of reducing the sample size to both choose good quantile levels and to estimate the within-class quantiles. Secondly, since the optimal quantile level is the same for every quantile, quantile-based classifiers is able to borrow information across all of the features to estimate the optimal quantile level, while composite quantile-based classifiers selects the quantile level one feature at-a-time. Thirdly, the freedom to select linear combination coefficients for the composite quantile-based classifiers method can result in additional variation. These points highlight the trade-off between the two methods. When the optimal quantile levels and discriminatory information are the same across the components then quantile-based classifiers perform better than CQC, and when this is not the case then CQC may perform better in some settings.

Simulation results for the block transformed Gaussian setting are shown in Table A.3 and Table A.4. In contrast to the previous scenarios, in this setting the optimal quantile levels and discriminatory information vary across the components, and in this setting CQC typically performs better than quantile-based classifiers. FANS and decision trees are the other best-performing classifiers in this scenario. The inclusion of decision trees suggests that there are some relatively discriminative features. When the quantity of the training data decreases to $n = 50$ we see a sharp rise in the misclassification rates for all methods. It is in this setting that the hybrid CQC and quantile-based classifiers do well. For example, when $n = 50$ and $p = 500$ with uncorrelated data, using the hybrid approach reduces the misclassification rate from 0.351 to 0.215. Both CQC and quantile-based classifiers perform well in this setting compared to other methods. We attribute this in part to the relative simplicity of the models in this data-sparse setting. Additionally, the reason that the hybrid approach works well here is that we may have a number of relatively discriminative features that the quantile-based methods approach of selecting quantile levels can effectively key in on, and then the additional linear combination step can help to deal with some of the differences in scale and discriminatory power between the blocks.

Simulation results for the beta distributed and gamma distributed data settings are shown in Table A.5 and Table A.6. In the case of the beta distributed data, CQC and decision trees performed the best. We see this as a sign of a few features having some degree of separability across the two classes. In this setting where the optimal quantile level varies across all of the features, we see that CQC having more flexibility to select quantile levels results in better performance as compared to quantile-based classifiers. We see similar results for the gamma distributed data, although in this case quantile-based classifiers perform about as well as CQC. We see this as suggesting that there happen to be a number of features with similar optimal quantile levels that quantile-based classifiers can select to achieve similar performance as the more flexible approach.

Simulation results for the Gaussian distributed data settings with partially equal mean vectors are shown in Table A.7. We note that due to the construction of this scenario, the marginal distributions for the even-numbered features are identical between the classes and will be deleted by marginal-based methods such as composite quantile-based classifiers, naive Bayes, and FANS. Now, the inverse of an AR1 matrix is a tridiagonal matrix, and the Bayes rule decision boundary is the set of solutions for x to $\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} = c$ for some constant c. But since $\boldsymbol{\Sigma}^{-1}$ is tridiagonal, it follows that the nonzero elements of $\boldsymbol{\mu}$ are spread out over the features that have a mean of 0 for $\boldsymbol{\mu}$, and as a result methods that delete these features will suffer from a loss of efficiency. On the other hand, since the decision rule boundary is linear, we expect that methods such as penalized logistic regression should be able to perform well.

In the simulation results, we see that the regular form of composite quantile-based classifiers does suffer from some loss of efficiency, compared to say penalized logistic regression. However, as expected, the augmented form of the classifiers does improve the efficiency of the classifier in the simulations. For example, when $n = 250$ and $p = 50$, we have a decrease in the misclassification rate from 0.122 to 0.039. The classification improves in nearly all of the combinations of n and p, but when the size of the training data is limited ($n = 50$), the improvement is fairly minimal. We expect that this is due to the variable selection performed by the regression step being overwhelmed by the doubling of the parameter space and the lack of training data.

Simulation results for the mixture Gaussian distributed data are shown in Table A.8. When the training data are more abundant ($n = 250$ and $n = 500$) then the multimodal composite quantile-based classifiers has either the lowest or nearly the lowest misclassification rate, with FANS a close competitor for some settings. For example, when $n = 250$ and $p = 250$, then the composite quantile-based classifiers have a misclassification rate of 0.083 compared to a next-best rate of 0.117 among other competitors. When $n = 500$ then composite quantile-based classifiers and FANS have similar misclassification rates across the various choices of p, with the two methods performing much better than any of the other classifiers. For example, when $n = 500$ and $p = 500$, then composite quantile-based classifiers and FANS have a misclassification rate of 0.060 and 0.056 respectively, while the next closest competitor has a misclassification rate of 0.191.

It is interesting to note that when the training data is less abundant ($n = 50$), quantile-based classifiers has the best empirical misclassification rate when $p = 250$ or $p = 500$, despite its inability to classify multiple intervals to a single class as is necessary to achieve the Bayes rule misclassification rate. We attribute this to the fact that when the sample size is small, quantile-based classifiers is better able to approximate its optimal classifier than other methods. However we observe that when the training data is more abundant, then as we would expect the misclassification rate has a higher lower bound than for other classifiers.

Simulation results comparing the model fitting algorithms described in Section 3.3 and Section 3.4 of the main paper are shown in Table A.9 and Table A.10. In each of these tables, we consider several different classifier model selection algorithms in addition to the usual approach. In the first and second classifiers, we simultaneously include the transformed data for each of the features across a grid of quantile levels. In the first classifier, we use a grid of quantile levels with values 0.01, 0.02, . . . , 0.99, and in the second classifier, we use a grid of quantile levels with values 0.05, 0.10, . . . , 0.95. These two classifiers are labeled as CQC simultaneous (0.01) and CQC simultaneous (0.05), in reference to their grid step size. For yet another classifier variant, we also use a quantile level grid with step size 0.05, but this time we use feature screening to reduce the number of features to twice the number of the original number of features. We label this variant as CQC simultaneous screen. We additionally include an augmented version that includes the original data for each of these classifier algorithms.

In Table A.9 we see that the usual CQC approach augmented with the original data performs the best in all of the scenarios. We expect the CQC models with the augmented data to perform better than the models without augmented data since the Bayes decision rule boundary is a linear boundary and the classifier constructed from the transformed data has a piecewise linear form. Among the CQC models that were augmented with the original data, we find that the classifiers that did not perform simultaneous selection performed better than those that did. We expect that this is because it is important to keep the original features in the model, and including the transformed features for all of the quantile levels makes it less likely that they will have much or any contribution to the prediction.

In Table A.10, we see that many of the various classifiers achieve the best classification rate for at least one of the simulation settings. In fact this is quite representative of our experience in comparing these various classifier algorithms over many different simulation studies. In short, the different algorithms seemed fairly stable in the sense that there wasn't usually a wide variability in the classification rates, and no one type of classifier was consistently better than the others.

A.3.4 Categorical and mixed data

In practice, one often encounters data that has variables with both continuous as well as categorical data. In problems such as these categorical data is typically reformulated into so-called dummy-coded variables. However, the quantile-based classifier is inherently ill-conditioned to handle such data. When categorical variables are present in the data, we propose including them in the classifier as untransformed dummy-coded variables.

A second type of data can also cause problems: one where the data is a mixture of point masses as well as continuous data. As an example, consider univariate data where the the quantiles of the underlying populations are given as shown in Figure A.4.

Figure A.4: Example mixture distribution with both categorical and continuous values

	Quantile levels											
	0.01	···	0.89	0.90	0.91	0.92	0.93	0.94	0.95	0.96	0.97	0.98	0.99
Class 0	0.00	···	0.00	0.02	0.14	0.23	0.33	0.42	0.55	0.68	0.87	1.05	1.41
Class 1	0.00	···	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.06	0.08	0.18	0.20

This example raises the question as to what the quantile distances would be for a fixed choice of quantile level. Suppose first that we consider a quantile level with nonzero quantiles for both classes, say 0.97. In this case, the quantile distances for the data in class 0 with value 0 are larger than the quantile distances for the data in class 1 with value 0. As a result, the quantile-based data will have spuriously introduced differences in the data for 90% of the data, when in fact there is no discriminative information in those quantiles. Conversely, suppose now that we select a quantile level with quantiles for both classes with value 0. Doing so has the effect of simply multiplying the nonzero values in the data by a constant factor (i.e. the chosen quantile level). While this is less disastrous then the previous scenario, we are not using the quantile-based data approach consistent with the rest of the paper.

What we propose to do for such a scenario is the following. When mixture data such as the example given above is present for a variable, we try to separate the parts of the data that belong to the point mass contribution to the mixture, and those that belong to the continuous part of the mixture. Then for the point mass data, we create a categorical variable that includes a reference category for the continuous part of the data, and for the continuous part we perform the quantile-based classification using only the reduced subset of the data.

Table A.1: Gaussian data, correlation coefficient = 0.8

	$n = 50$	$n = 250$	$n = 500$
	$p = 50$		
CQC	0.402 (0.02)	0.361 (0.02)	0.358 (0.01)
CQC multimodal	0.401 (0.02)	0.362 (0.02)	0.357 (0.01)
CQC augmented	0.391 (0.02)	0.353 (0.02)	0.349 (0.01)
CQC multimodal augmented	0.390 (0.02)	0.356 (0.02)	0.351 (0.01)
Quantile-based classifiers	0.414 (0.04)	0.359 (0.01)	0.349 (0.01)
FANS	0.446 (0.04)	0.352 (0.01)	0.362 (0.01)
FANS2	0.439 (0.05)	0.356 (0.02)	0.352 (0.01)
Penalized logistic regression	0.396 (0.03)	0.360 (0.02)	0.354 (0.01)
Support vector machine	0.389 (0.03)	0.349 (0.02)	0.344 (0.01)
k-nearest neighbor	0.429 (0.02)	0.401 (0.04)	0.387 (0.01)
Naive Bayes	**0.375 (0.03)**	0.339 (0.01)	0.339 (0.01)
Nearest shrunken centroids	0.381 (0.03)	0.341 (0.01)	0.340 (0.02)
Penalized LDA	0.376 (0.05)	**0.336 (0.02)**	**0.337 (0.01)**
Decision trees	0.433 (0.03)	0.408 (0.02)	0.403 (0.01)
	$p = 250$		
CQC	0.437 (0.02)	0.380 (0.02)	0.370 (0.01)
CQC multimodal	0.428 (0.03)	0.385 (0.01)	0.368 (0.02)
CQC augmented	0.429 (0.02)	0.379 (0.02)	0.368 (0.01)
CQC multimodal augmented	0.426 (0.02)	0.384 (0.02)	0.365 (0.02)
Quantile-based classifiers	0.428 (0.03)	0.394 (0.03)	0.362 (0.02)
FANS	0.496 (0.01)	0.409 (0.02)	0.375 (0.01)
FANS2	0.474 (0.03)	0.398 (0.02)	0.360 (0.01)
Penalized logistic regression	0.446 (0.02)	0.395 (0.02)	0.369 (0.01)
Support vector machine	0.403 (0.03)	0.370 (0.02)	0.349 (0.02)
k-nearest neighbor	0.430 (0.04)	0.392 (0.02)	0.384 (0.01)
Naive Bayes	0.401 (0.02)	0.349 (0.01)	0.342 (0.02)
Nearest shrunken centroids	0.379 (0.02)	**0.344 (0.01)**	**0.336 (0.01)**
Penalized LDA	**0.374 (0.02)**	**0.344 (0.01)**	0.337 (0.01)
Decision trees	0.459 (0.02)	0.447 (0.01)	0.438 (0.03)
	$p = 500$		
CQC	0.427 (0.03)	0.394 (0.02)	0.369 (0.02)
CQC multimodal	0.437 (0.03)	0.386 (0.02)	0.379 (0.02)
CQC augmented	0.442 (0.04)	0.388 (0.02)	0.362 (0.01)
CQC multimodal augmented	0.443 (0.04)	0.379 (0.20)	0.362 (0.02)
Quantile-based classifiers	0.402 (0.02)	0.402 (0.03)	0.395 (0.01)
FANS	0.483 (0.03)	0.428 (0.03)	0.375 (0.02)
FANS2	0.468 (0.04)	0.411 (0.04)	0.368 (0.02)
Penalized logistic regression	0.441 (0.04)	0.387 (0.03)	0.360 (0.02)
Support vector machine	0.385 (0.03)	0.365 (0.03)	0.349 (0.01)
k-nearest neighbor	0.403 (0.04)	0.386 (0.04)	0.386 (0.03)
Naive Bayes	0.391 (0.03)	0.360 (0.02)	0.338 (0.01)
Nearest shrunken centroids	**0.349 (0.03)**	0.349 (0.04)	**0.331 (0.01)**
Penalized LDA	0.379 (0.05)	**0.346 (0.02)**	0.336 (0.01)
Decision trees	0.480 (0.03)	0.444 (0.03)	0.434 (0.02)

Table A.2: Exponentiated Gaussian data, correlation coefficient = 0.8

	$n = 50$	$n = 250$	$n = 500$
	$p = 50$		
CQC	0.223 (0.02)	0.120 (0.01)	0.098 (0.01)
CQC multimodal	0.247 (0.02)	0.124 (0.02)	0.100 (0.01)
CQC augmented	0.232 (0.02)	0.132 (0.01)	0.108 (0.01)
CQC multimodal augmented	0.265 (0.02)	0.133 (0.01)	0.109 (0.01)
Quantile-based classifiers	**0.133 (0.03)**	**0.064 (0.01)**	**0.061 (0.01)**
FANS	0.388 (0.06)	0.232 (0.03)	0.177 (0.02)
FANS2	0.405 (0.04)	0.235 (0.02)	0.184 (0.02)
Penalized logistic regression	0.430 (0.02)	0.426 (0.01)	0.430 (0.01)
Support vector machine	0.417 (0.03)	0.362 (0.02)	0.355 (0.02)
k-nearest neighbor	0.446 (0.02)	0.412 (0.02)	0.416 (0.02)
Naive Bayes	0.453 (0.02)	0.445 (0.03)	0.436 (0.03)
Nearest shrunken centroids	0.432 (0.02)	0.418 (0.03)	0.414 (0.02)
Penalized LDA	0.432 (0.04)	0.412 (0.01)	0.405 (0.01)
Decision trees	0.327 (0.02)	0.201 (0.03)	0.155 (0.02)
	$p = 250$		
CQC	0.242 (0.03)	0.164 (0.02)	0.128 (0.01)
CQC multimodal	0.335 (0.04)	0.164 (0.02)	0.132 (0.01)
CQC augmented	0.305 (0.03)	0.167 (0.02)	0.134 (0.01)
CQC multimodal augmented	0.385 (0.04)	0.172 (0.01)	0.109 (0.01)
Quantile-based classifiers	**0.157 (0.03)**	**0.068 (0.01)**	**0.060 (0.01)**
FANS	0.457 (0.05)	0.265 (0.02)	0.197 (0.02)
FANS2	0.463 (0.05)	0.271 (0.02)	0.197 (0.01)
Penalized logistic regression	0.489 (0.02)	0.458 (0.03)	0.435 (0.02)
Support vector machine	0.450 (0.02)	0.433 (0.02)	0.404 (0.02)
k-nearest neighbor	0.448 (0.02)	0.417 (0.03)	0.403 (0.02)
Naive Bayes	0.465 (0.01)	0.461 (0.03)	0.450 (0.03)
Nearest shrunken centroids	0.421 (0.02)	0.434 (0.02)	0.414 (0.02)
Penalized LDA	0.464 (0.03)	0.438 (0.04)	0.407 (0.01)
Decision trees	0.416 (0.05)	0.208 (0.03)	0.144 (0.02)
	$p = 500$		
CQC	0.268 (0.02)	0.160 (0.01)	0.135 (0.01)
CQC multimodal	0.370 (0.05)	0.164 (0.01)	0.140 (0.01)
CQC augmented	0.344 (0.05)	0.172 (0.02)	0.138 (0.01)
CQC multimodal augmented	0.399 (0.03)	0.168 (0.01)	0.139 (0.01)
Quantile-based classifiers	**0.169 (0.03)**	**0.065 (0.01)**	**0.060 (0.01)**
FANS	0.453 (0.05)	0.268 (0.03)	0.205 (0.02)
FANS2	0.476 (0.04)	0.271 (0.03)	0.206 (0.02)
Penalized logistic regression	0.490 (0.01)	0.458 (0.02)	0.461 (0.02)
Support vector machine	0.459 (0.03)	0.453 (0.02)	0.441 (0.02)
k-nearest neighbor	0.438 (0.04)	0.428 (0.02)	0.415 (0.02)
Naive Bayes	0.471 (0.02)	0.458 (0.02)	0.456 (0.03)
Nearest shrunken centroids	0.427 (0.02)	0.424 (0.02)	0.426 (0.02)
Penalized LDA	0.459 (0.03)	0.436 (0.02)	0.426 (0.02)
Decision trees	0.412 (0.07)	0.196 (0.03)	0.156 (0.02)

Table A.3: Block transformed data, correlation coefficient = 0

	$n = 50$	$n = 250$	$n = 500$
	$p = 50$		
CQC	0.181 (0.03)	**0.043 (0.01)**	0.029 (0.01)
CQC multimodal	0.217 (0.06)	0.045 (0.01)	0.029 (0.01)
CQC augmented	**0.176 (0.05)**	0.048 (0.01)	0.030 (0.01)
CQC multimodal augmented	0.225 (0.05)	0.046 (0.01)	0.031 (0.01)
Quantile-based classifiers	0.223 (0.03)	0.112 (0.02)	0.085 (0.01)
FANS	0.320 (0.08)	0.068 (0.01)	**0.027 (0.01)**
FANS2	0.323 (0.06)	0.063 (0.01)	0.028 (0.01)
Penalized logistic regression	0.431 (0.03)	0.322 (0.02)	0.291 (0.02)
Support vector machine	0.386 (0.02)	0.306 (0.02)	0.284 (0.01)
k-nearest neighbor	0.464 (0.02)	0.448 (0.02)	0.431 (0.02)
Naive Bayes	0.452 (0.01)	0.396 (0.02)	0.374 (0.02)
Nearest shrunken centroids	0.422 (0.05)	0.339 (0.01)	0.330 (0.02)
Penalized LDA	0.374 (0.03)	0.301 (0.02)	0.283 (0.01)
Decision trees	0.374 (0.08)	0.078 (0.01)	0.040 (0.01)
	$p = 250$		
CQC	**0.222 (0.08)**	**0.062 (0.01)**	0.045 (0.01)
CQC multimodal	0.286 (0.05)	**0.062 (0.01)**	0.047 (0.01)
CQC augmented	0.234 (0.08)	0.063 (0.01)	0.047 (0.01)
CQC multimodal augmented	0.316 (0.06)	0.067 (0.01)	0.049 (0.01)
Quantile-based classifiers	0.234 (0.04)	0.115 (0.01)	0.089 (0.01)
FANS	0.374 (0.07)	0.083 (0.01)	**0.036 (0.01)**
FANS2	0.398 (0.08)	0.080 (0.01)	0.038 (0.01)
Penalized logistic regression	0.465 (0.03)	0.382 (0.03)	0.343 (0.02)
Support vector machine	0.424 (0.02)	0.382 (0.01)	0.344 (0.01)
k-nearest neighbor	0.470 (0.03)	0.446 (0.03)	0.451 (0.02)
Naive Bayes	0.472 (0.01)	0.426 (0.02)	0.408 (0.01)
Nearest shrunken centroids	0.403 (0.05)	0.318 (0.02)	0.303 (0.02)
Penalized LDA	0.435 (0.03)	0.372 (0.02)	0.340 (0.01)
Decision trees	0.354 (0.07)	0.072 (0.02)	0.038 (0.01)
	$p = 500$		
CQC	**0.215 (0.04)**	**0.065 (0.01)**	0.046 (0.01)
CQC multimodal	0.332 (0.06)	**0.065 (0.01)**	0.047 (0.01)
CQC augmented	0.267 (0.07)	0.070 (0.01)	0.050 (0.01)
CQC multimodal augmented	0.390 (0.04)	0.070 (0.01)	0.050 (0.01)
Quantile-based classifiers	0.244 (0.04)	0.131 (0.02)	0.093 (0.02)
FANS	0.404 (0.07)	0.102 (0.02)	**0.033 (0.01)**
FANS2	0.422 (0.07)	0.094 (0.02)	**0.033 (0.01)**
Penalized logistic regression	0.472 (0.03)	0.414 (0.04)	0.342 (0.01)
Support vector machine	0.451 (0.02)	0.400 (0.01)	0.367 (0.01)
k-nearest neighbor	0.464 (0.03)	0.439 (0.02)	0.434 (0.02)
Naive Bayes	0.474 (0.02)	0.440 (0.02)	0.420 (0.02)
Nearest shrunken centroids	0.406 (0.04)	0.318 (0.01)	0.300 (0.01)
Penalized LDA	0.470 (0.03)	0.392 (0.02)	0.358 (0.02)
Decision trees	0.406 (0.09)	0.076 (0.03)	0.043 (0.02)

Table A.4: Block transformed data, correlation coefficient = 0.8

	$n = 50$	$n = 250$	$n = 500$
	$p = 50$		
CQC	**0.224 (0.04)**	**0.081 (0.01)**	**0.064 (0.01)**
CQC multimodal	**0.224 (0.05)**	0.084 (0.01)	0.068 (0.01)
CQC augmented	0.234 (0.06)	0.086 (0.01)	0.066 (0.01)
CQC multimodal augmented	0.234 (0.05)	0.086 (0.01)	0.068 (0.01)
Quantile-based classifiers	0.280 (0.08)	0.134 (0.02)	0.107 (0.01)
FANS	0.302 (0.06)	0.115 (0.02)	0.076 (0.01)
FANS2	0.308 (0.06)	0.117 (0.01)	0.076 (0.01)
Penalized logistic regression	0.443 (0.03)	0.409 (0.02)	0.396 (0.02)
Support vector machine	0.420 (0.03)	0.384 (0.02)	0.375 (0.03)
k-nearest neighbor	0.480 (0.03)	0.450 (0.02)	0.444 (0.02)
Naive Bayes	0.454 (0.04)	0.420 (0.02)	0.410 (0.02)
Nearest shrunken centroids	0.462 (0.04)	0.422 (0.01)	0.416 (0.03)
Penalized LDA	0.430 (0.05)	0.388 (0.02)	0.383 (0.02)
Decision trees	0.313 (0.07)	0.115 (0.02)	0.076 (0.01)
	$p = 250$		
CQC	**0.228 (0.03)**	**0.115 (0.01)**	0.091 (0.00)
CQC multimodal	0.339 (0.06)	**0.115 (0.01)**	0.094 (0.01)
CQC augmented	0.293 (0.05)	0.118 (0.01)	0.092 (0.00)
CQC multimodal augmented	0.383 (0.06)	0.116 (0.01)	0.091 (0.01)
Quantile-based classifiers	0.263 (0.02)	0.154 (0.02)	0.126 (0.03)
FANS	0.428 (0.08)	0.128 (0.01)	**0.079 (0.01)**
FANS2	0.402 (0.07)	0.139 (0.01)	0.085 (0.01)
Penalized logistic regression	0.477 (0.03)	0.432 (0.03)	0.421 (0.02)
Support vector machine	0.452 (0.02)	0.428 (0.03)	0.416 (0.02)
k-nearest neighbor	0.463 (0.02)	0.450 (0.02)	0.452 (0.03)
Naive Bayes	0.466 (0.02)	0.436 (0.02)	0.435 (0.02)
Nearest shrunken centroids	0.434 (0.02)	0.388 (0.02)	0.392 (0.02)
Penalized LDA	0.460 (0.03)	0.403 (0.02)	0.397 (0.02)
Decision trees	0.443 (0.07)	0.118 (0.02)	0.081 (0.01)
	$p = 500$		
CQC	**0.238 (0.04)**	0.125 (0.01)	0.090 (0.01)
CQC multimodal	0.379 (0.06)	**0.123 (0.01)**	0.091 (0.01)
CQC augmented	0.312 (0.06)	**0.123 (0.01)**	0.090 (0.01)
CQC multimodal augmented	0.404 (0.05)	0.127 (0.01)	0.090 (0.01)
Quantile-based classifiers	0.277 (0.03)	0.158 (0.02)	0.115 (0.03)
FANS	0.429 (0.07)	0.146 (0.02)	0.092 (0.01)
FANS2	0.415 (0.07)	0.148 (0.02)	0.095 (0.01)
Penalized logistic regression	0.480 (0.02)	0.444 (0.02)	0.430 (0.02)
Support vector machine	0.460 (0.02)	0.436 (0.02)	0.427 (0.02)
k-nearest neighbor	0.487 (0.01)	0.457 (0.01)	0.459 (0.02)
Naive Bayes	0.480 (0.03)	0.450 (0.02)	0.436 (0.01)
Nearest shrunken centroids	0.423 (0.03)	0.398 (0.03)	0.397 (0.02)
Penalized LDA	0.465 (0.04)	0.423 (0.02)	0.411 (0.01)
Decision trees	0.421 (0.10)	0.146 (0.03)	**0.084 (0.02)**

Table A.5: Beta distributed data

	$n = 50$	$n = 250$	$n = 500$
	$p = 50$		
CQC	0.060 (0.04)	0.016 (0.01)	**0.009 (0.01)**
CQC multimodal	0.064 (0.04)	0.015 (0.01)	0.010 (0.01)
CQC augmented	0.068 (0.03)	0.015 (0.01)	0.010 (0.01)
CQC multimodal augmented	0.077 (0.03)	**0.013 (0.01)**	**0.009 (0.00)**
Quantile-based classifiers	0.158 (0.03)	0.031 (0.01)	0.019 (0.01)
FANS	**0.053 (0.03)**	0.025 (0.02)	0.015 (0.00)
FANS2	0.072 (0.05)	0.023 (0.01)	0.016 (0.00)
Penalized logistic regression	0.420 (0.17)	0.511 (0.02)	0.501 (0.02)
Support vector machine	0.445 (0.10)	0.475 (0.03)	0.438 (0.02)
k-nearest neighbor	0.505 (0.02)	0.491 (0.01)	0.484 (0.01)
Naive Bayes	0.195 (0.16)	0.332 (0.13)	0.327 (0.11)
Nearest shrunken centroids	0.498 (0.01)	0.512 (0.01)	0.495 (0.02)
Penalized LDA	0.420 (0.17)	0.503 (0.01)	0.507 (0.01)
Decision trees	0.071 (0.03)	0.029 (0.01)	0.019 (0.00)
	$p = 250$		
CQC	**0.049 (0.02)**	**0.011 (0.00)**	0.012 (0.00)
CQC multimodal	0.050 (0.02)	0.015 (0.01)	0.013 (0.01)
CQC augmented	0.063 (0.03)	**0.011 (0.01)**	**0.007 (0.00)**
CQC multimodal augmented	0.068 (0.04)	0.014 (0.01)	0.010 (0.01)
Quantile-based classifiers	0.301 (0.02)	0.173 (0.03)	0.108 (0.01)
FANS	0.059 (0.04)	0.024 (0.01)	0.018 (0.01)
FANS2	0.059 (0.04)	0.026 (0.01)	0.017 (0.01)
Penalized logistic regression	0.464 (0.11)	0.495 (0.02)	0.505 (0.02)
Support vector machine	0.499 (0.02)	0.500 (0.02)	0.494 (0.02)
k-nearest neighbor	0.507 (0.02)	0.490 (0.02)	0.490 (0.02)
Naive Bayes	0.434 (0.06)	0.331 (0.18)	0.413 (0.10)
Nearest shrunken centroids	0.496 (0.01)	0.494 (0.02)	0.504 (0.01)
Penalized LDA	0.494 (0.01)	0.500 (0.01)	0.501 (0.02)
Decision trees	0.076 (0.06)	0.022 (0.01)	0.016 (0.00)
	$p = 500$		
CQC	0.075 (0.03)	**0.019 (0.01)**	**0.011 (0.01)**
CQC multimodal	0.076 (0.04)	**0.019 (0.01)**	0.014 (0.01)
CQC augmented	0.068 (0.03)	0.021 (0.01)	0.012 (0.00)
CQC multimodal augmented	0.082 (0.03)	0.020 (0.01)	0.014 (0.01)
Quantile-based classifiers	0.356 (0.02)	0.228 (0.03)	0.164 (0.02)
FANS	0.073 (0.07)	0.028 (0.01)	0.020 (0.01)
FANS2	0.072 (0.05)	0.031 (0.01)	0.022 (0.01)
Penalized logistic regression	0.460 (0.13)	0.485 (0.02)	0.506 (0.01)
Support vector machine	0.497 (0.02)	0.501 (0.02)	0.506 (0.02)
k-nearest neighbor	0.499 (0.02)	0.498 (0.01)	0.495 (0.02)
Naive Bayes	0.462 (0.03)	0.441 (0.05)	0.459 (0.06)
Nearest shrunken centroids	0.497 (0.01)	0.492 (0.02)	0.500 (0.01)
Penalized LDA	0.491 (0.02)	0.498 (0.02)	0.498 (0.01)
Decision trees	**0.060 (0.03)**	0.030 (0.01)	0.013 (0.01)

Table A.6: Gamma distributed data

	$n = 50$	$n = 250$	$n = 500$
	$p = 50$		
CQC	**0.128 (0.03)**	**0.052 (0.01)**	0.035 (0.01)
CQC multimodal	0.159 (0.04)	**0.052 (0.01)**	0.037 (0.01)
CQC augmented	0.153 (0.04)	0.054 (0.01)	**0.034 (0.01)**
CQC multimodal augmented	0.191 (0.05)	**0.052 (0.01)**	**0.034 (0.01)**
Quantile-based classifiers	0.156 (0.01)	0.062 (0.02)	0.045 (0.01)
FANS	0.389 (0.10)	0.107 (0.02)	0.069 (0.01)
FANS2	0.363 (0.09)	0.116 (0.02)	0.073 (0.01)
Penalized logistic regression	0.493 (0.01)	0.500 (0.02)	0.501 (0.02)
Support vector machine	0.492 (0.01)	0.462 (0.02)	0.424 (0.01)
k-nearest neighbor	0.495 (0.01)	0.489 (0.02)	0.486 (0.02)
Naive Bayes	0.424 (0.02)	0.369 (0.02)	0.341 (0.02)
Nearest shrunken centroids	0.506 (0.02)	0.506 (0.02)	0.503 (0.01)
Penalized LDA	0.498 (0.00)	0.506 (0.02)	0.501 (0.01)
Decision trees	0.272 (0.11)	0.060 (0.01)	0.038 (0.01)
	$p = 250$		
CQC	**0.228 (0.12)**	**0.070 (0.02)**	0.041 (0.01)
CQC multimodal	0.242 (0.07)	0.073 (0.02)	0.044 (0.01)
CQC augmented	0.272 (0.11)	0.083 (0.03)	0.046 (0.01)
CQC multimodal augmented	0.302 (0.05)	0.086 (0.02)	0.048 (0.01)
Quantile-based classifiers	0.246 (0.13)	**0.070 (0.02)**	0.055 (0.02)
FANS	0.392 (0.09)	0.146 (0.03)	0.093 (0.03)
FANS2	0.413 (0.07)	0.148 (0.04)	0.113 (0.05)
Penalized logistic regression	0.503 (0.01)	0.503 (0.01)	0.502 (0.01)
Support vector machine	0.495 (0.01)	0.501 (0.01)	0.496 (0.01)
k-nearest neighbor	0.492 (0.01)	0.489 (0.02)	0.481 (0.01)
Naive Bayes	0.501 (0.01)	0.494 (0.01)	0.491 (0.01)
Nearest shrunken centroids	0.493 (0.01)	0.501 (0.01)	0.492 (0.02)
Penalized LDA	0.498 (0.01)	0.500 (0.01)	0.500 (0.00)
Decision trees	0.296 (0.09)	0.075 (0.03)	**0.040 (0.01)**
	$p = 500$		
CQC	0.211 (0.05)	**0.070 (0.01)**	0.045 (0.01)
CQC multimodal	0.359 (0.11)	0.076 (0.01)	0.045 (0.01)
CQC augmented	0.335 (0.08)	0.081 (0.01)	0.049 (0.01)
CQC multimodal augmented	0.406 (0.08)	0.083 (0.01)	0.055 (0.02)
Quantile-based classifiers	**0.192 (0.03)**	0.089 (0.02)	0.062 (0.01)
FANS	0.374 (0.11)	0.128 (0.02)	0.083 (0.01)
FANS2	0.424 (0.09)	0.162 (0.05)	0.091 (0.02)
Penalized logistic regression	0.503 (0.01)	0.503 (0.02)	0.502 (0.01)
Support vector machine	0.501 (0.00)	0.496 (0.01)	0.506 (0.01)
k-nearest neighbor	0.499 (0.02)	0.490 (0.02)	0.496 (0.02)
Naive Bayes	0.498 (0.01)	0.494 (0.01)	0.492 (0.01)
Nearest shrunken centroids	0.504 (0.01)	0.504 (0.02)	0.504 (0.01)
Penalized LDA	0.496 (0.01)	0.501 (0.01)	0.504 (0.01)
Decision trees	0.341 (0.12)	0.110 (0.13)	**0.039 (0.01)**

Table A.7: Gaussian data with partially equivalent marginals, correlation coefficient = 0.8

	$n = 50$	$n = 250$	$n = 500$
	$p = 50$		
CQC	0.302 (0.03)	0.122 (0.02)	0.076 (0.01)
CQC multimodal	0.301 (0.02)	0.127 (0.02)	0.080 (0.01)
CQC augmented	0.276 (0.03)	0.034 (0.01)	0.014 (0.00)
CQC multimodal augmented	0.262 (0.02)	0.039 (0.01)	0.013 (0.00)
Quantile-based classifiers	0.354 (0.04)	0.302 (0.03)	0.281 (0.02)
FANS	0.419 (0.03)	0.094 (0.01)	0.049 (0.01)
FANS2	0.429 (0.05)	0.110 (0.02)	0.051 (0.01)
Penalized logistic regression	0.206 (0.05)	0.023 (0.00)	0.012 (0.00)
Support vector machine	0.**097 (0.04)**	**0.012 (0.00)**	**0.008 (0.00)**
k-nearest neighbor	0.340 (0.03)	0.263 (0.02)	0.224 (0.02)
Naive Bayes	0.321 (0.03)	0.277 (0.02)	0.262 (0.02)
Nearest shrunken centroids	0.312 (0.03)	0.269 (0.02)	0.259 (0.01)
Penalized LDA	0.310 (0.02)	0.275 (0.02)	0.265 (0.02)
Decision trees	0.415 (0.02)	0.375 (0.02)	0.345 (0.02)
	$p = 250$		
CQC	0.384 (0.04)	0.281 (0.02)	0.188 (0.02)
CQC multimodal	0.395 (0.03)	0.276 (0.02)	0.186 (0.02)
CQC augmented	0.376 (0.04)	0.214 (0.03)	0.043 (0.01)
CQC multimodal augmented	0.371 (0.03)	0.209 (0.02)	0.045 (0.01)
Quantile-based classifiers	0.366 (0.02)	0.335 (0.04)	0.321 (0.05)
FANS	0.444 (0.03)	0.117 (0.02)	0.056 (0.01)
FANS2	0.472 (0.06)	0.127 (0.02)	0.056 (0.01)
Penalized logistic regression	0.384 (0.04)	**0.064 (0.01)**	**0.020 (0.00)**
Support vector machine	0.334 (0.03)	0.135 (0.01)	0.047 (0.01)
k-nearest neighbor	0.347 (0.03)	0.265 (0.03)	0.225 (0.02)
Naive Bayes	0.368 (0.02)	0.296 (0.01)	0.282 (0.02)
Nearest shrunken centroids	**0.316 (0.04)**	0.260 (0.01)	0.264 (0.02)
Penalized LDA	0.340 (0.02)	0.276 (0.02)	0.272 (0.01)
Decision trees	0.451 (0.04)	0.417 (0.02)	0.394 (0.02)
	$p = 500$		
CQC	0.430 (0.03)	0.311 (0.02)	0.241 (0.03)
CQC multimodal	0.413 (0.03)	0.308 (0.02)	0.245 (0.02)
CQC augmented	0.407 (0.04)	0.278 (0.02)	0.095 (0.02)
CQC multimodal augmented	0.419 (0.05)	0.279 (0.03)	0.094 (0.03)
Quantile-based classifiers	0.370 (0.02)	0.375 (0.02)	0.366 (0.02)
FANS	0.488 (0.02)	0.140 (0.01)	0.054 (0.01)
FANS2	0.486 (0.02)	0.160 (0.01)	0.058 (0.01)
Penalized logistic regression	0.402 (0.03)	**0.100 (0.03)**	**0.023 (0.01)**
Support vector machine	0.367 (0.02)	0.242 (0.02)	0.156 (0.02)
k-nearest neighbor	0.350 (0.02)	0.264 (0.02)	0.232 (0.02)
Naive Bayes	0.392 (0.03)	0.314 (0.02)	0.302 (0.02)
Nearest shrunken centroids	**0.338 (0.05)**	0.259 (0.04)	0.263 (0.02)
Penalized LDA	0.375 (0.05)	0.288 (0.02)	0.288 (0.02)
Decision trees	0.439 (0.04)	0.409 (0.03)	0.397 (0.02)

Table A.8: Mixture Gaussian data, correlation coefficient = 0

	$n = 50$	$n = 250$	$n = 500$
	$p = 50$		
CQC	0.341 (0.03)	0.188 (0.01)	0.164 (0.01)
CQC multimodal	**0.261 (0.03)**	**0.064 (0.01)**	**0.046 (0.00)**
CQC augmented	0.411 (0.03)	0.195 (0.02)	0.175 (0.03)
CQC multimodal augmented	0.334 (0.03)	0.076 (0.01)	0.053 (0.01)
Quantile-based classifiers	0.284 (0.05)	0.160 (0.02)	0.146 (0.01)
FANS	0.419 (0.03)	0.094 (0.01)	0.049 (0.01)
FANS2	0.429 (0.05)	0.110 (0.02)	0.051 (0.01)
Penalized logistic regression	0.494 (0.02)	0.503 (0.01)	0.502 (0.01)
Support vector machine	0.366 (0.02)	0.219 (0.02)	0.189 (0.02)
k-nearest neighbor	0.483 (0.01)	0.485 (0.01)	0.484 (0.01)
Naive Bayes	0.328 (0.02)	0.209 (0.02)	0.186 (0.02)
Nearest shrunken centroids	0.495 (0.01)	0.499 (0.02)	0.499 (0.02)
Penalized LDA	0.494 (0.01)	0.491 (0.01)	0.496 (0.01)
Decision trees	0.464 (0.03)	0.360 (0.04)	0.324 (0.04)
	$p = 250$		
CQC	0.464 (0.02)	0.234 (0.02)	0.187 (0.02)
CQC multimodal	0.400 (0.05)	**0.083 (0.01)**	**0.054 (0.01)**
CQC augmented	0.481 (0.02)	0.264 (0.03)	0.195 (0.03)
CQC multimodal augmented	0.442 (0.03)	0.096 (0.01)	0.061 (0.01)
Quantile-based classifiers	**0.359 (0.03)**	0.192 (0.02)	0.158 (0.01)
FANS	0.444 (0.03)	0.117 (0.02)	0.056 (0.01)
FANS2	0.472 (0.06)	0.127 (0.02)	0.057 (0.01)
Penalized logistic regression	0.499 (0.01)	0.495 (0.02)	0.499 (0.02)
Support vector machine	0.504 (0.01)	0.417 (0.05)	0.374 (0.02)
k-nearest neighbor	0.488 (0.01)	0.481 (0.01)	0.487 (0.01)
Naive Bayes	0.431 (0.02)	0.318 (0.03)	0.270 (0.02)
Nearest shrunken centroids	0.502 (0.01)	0.492 (0.02)	0.496 (0.01)
Penalized LDA	0.506 (0.01)	0.497 (0.02)	0.506 (0.02)
Decision trees	0.491 (0.02)	0.398 (0.06)	0.334 (0.04)
	$p = 500$		
CQC	0.483 (0.02)	0.259 (0.02)	0.211 (0.02)
CQC multimodal	0.437 (0.03)	**0.105 (0.01)**	0.060 (0.01)
CQC augmented	0.493 (0.02)	0.307 (0.02)	0.231 (0.02)
CQC multimodal augmented	0.471 (0.03)	0.124 (0.02)	0.064 (0.01)
Quantile-based classifiers	**0.370 (0.03)**	0.218 (0.02)	0.191 (0.02)
FANS	0.488 (0.02)	0.140 (0.01)	**0.056 (0.01)**
FANS2	0.486 (0.02)	0.160 (0.01)	0.058 (0.01)
Penalized logistic regression	0.501 (0.02)	0.500 (0.02)	0.494 (0.02)
Support vector machine	0.492 (0.02)	0.489 (0.01)	0.479 (0.02)
k-nearest neighbor	0.488 (0.01)	0.486 (0.01)	0.488 (0.01)
Naive Bayes	0.445 (0.02)	0.366 (0.01)	0.330 (0.02)
Nearest shrunken centroids	0.497 (0.02)	0.502 (0.01)	0.498 (0.01)
Penalized LDA	0.498 (0.01)	0.500 (0.00)	0.497 (0.01)
Decision trees	0.499 (0.02)	0.424 (0.04)	0.327 (0.04)

Table A.9: Model fitting comparison: Gaussian data, correlation coefficient = 0

	$n = 50$	$n = 250$	$n = 500$
	$p = 50$		
CQC	0.286 (0.03)	0.198 (0.02)	0.160 (0.01)
CQC augmented	**0.263 (0.05)**	**0.161 (0.02)**	**0.138 (0.01)**
CQC simultaneous (0.01)	0.338 (0.05)	0.203 (0.01)	0.169 (0.02)
CQC simultaneous (0.05)	0.329 (0.05)	0.202 (0.01)	0.166 (0.01)
CQC simultaneous screen	0.340 (0.04)	0.204 (0.01)	0.163 (0.01)
CQC simultaneous augmented (0.01)	0.330 (0.05)	0.203 (0.01)	0.167 (0.01)
CQC simultaneous augmented (0.05)	0.338 (0.06)	0.199 (0.01)	0.167 (0.01)
CQC simultaneous screen augmented	0.336 (0.06)	0.205 (0.01)	0.168 (0.01)
	$p = 250$		
CQC	0.391 (0.03)	0.243 (0.02)	0.196 (0.02)
CQC augmented	**0.355 (0.03)**	**0.199 (0.02)**	**0.158 (0.01)**
CQC simultaneous (0.01)	0.399 (0.03)	0.274 (0.03)	0.202 (0.02)
CQC simultaneous (0.05)	0.402 (0.03)	0.257 (0.02)	0.201 (0.02)
CQC simultaneous screen	0.413 (0.04)	0.251 (0.01)	0.199 (0.02)
CQC simultaneous augmented (0.01)	0.404 (0.02)	0.251 (0.02)	0.183 (0.02)
CQC simultaneous augmented (0.05)	0.395 (0.02)	0.251 (0.02)	0.183 (0.02)
CQC simultaneous screen augmented	0.405 (0.04)	0.246 (0.02)	0.186 (0.02)
	$p = 500$		
CQC	0.423 (0.03)	0.280 (0.02)	0.215 (0.02)
CQC augmented	**0.405 (0.04)**	**0.222 (0.02)**	**0.175 (0.01)**
CQC simultaneous (0.01)	0.436 (0.04)	0.289 (0.03)	0.220 (0.01)
CQC simultaneous (0.05)	0.445 (0.03)	0.274 (0.03)	0.221 (0.01)
CQC simultaneous screen	0.443 (0.04)	0.281 (0.04)	0.221 (0.01)
CQC simultaneous augmented (0.01)	0.439 (0.05)	0.275 (0.04)	0.214 (0.01)
CQC simultaneous augmented (0.05)	0.448 (0.03)	0.265 (0.03)	0.205 (0.01)
CQC simultaneous screen augmented	0.447 (0.04)	0.258 (0.03)	0.205 (0.01)

Table A.10: Model fitting comparison: Exponentiated Gaussian data, correlation coefficient = 0

	$n = 50$	$n = 250$	$n = 500$
	$p = 50$		
CQC	0.174 (0.04)	0.030 (0.01)	**0.012 (0.00)**
CQC augmented	0.206 (0.03)	**0.031 (0.01)**	**0.012 (0.00)**
CQC simultaneous (0.01)	**0.159 (0.03)**	0.038 (0.01)	0.013 (0.00)
CQC simultaneous (0.05)	0.165 (0.04)	0.033 (0.01)	0.013 (0.00)
CQC simultaneous screen	0.160 (0.04)	0.035 (0.01)	0.013 (0.00)
CQC simultaneous augmented (0.01)	0.167 (0.04)	0.040 (0.01)	0.014 (0.00)
CQC simultaneous augmented (0.05)	0.166 (0.04)	0.034 (0.01)	0.014 (0.00)
CQC simultaneous screen augmented	0.173 (0.04)	0.033 (0.01)	0.013 (0.00)
	$p = 250$		
CQC	**0.198 (0.03)**	0.031 (0.01)	0.017 (0.00)
CQC augmented	0.222 (0.04)	0.038 (0.01)	0.016 (0.00)
CQC simultaneous (0.01)	0.215 (0.04)	0.033 (0.01)	**0.014 (0.00)**
CQC simultaneous (0.05)	0.218 (0.04)	0.031 (0.01)	0.015 (0.00)
CQC simultaneous screen	0.218 (0.05)	**0.028 (0.00)**	**0.014 (0.00)**
CQC simultaneous augmented (0.01)	0.199 (0.03)	0.031 (0.01)	0.015 (0.00)
CQC simultaneous augmented (0.05)	0.213 (0.08)	0.030 (0.00)	0.015 (0.00)
CQC simultaneous screen augmented	0.200 (0.03)	0.031 (0.00)	0.015 (0.00)
	$p = 500$		
CQC	**0.225 (0.03)**	0.036 (0.01)	**0.016 (0.00)**
CQC augmented	0.292 (0.03)	0.043 (0.01)	0.017 (0.01)
CQC simultaneous (0.01)	0.232 (0.07)	0.032 (0.01)	0.017 (0.00)
CQC simultaneous (0.05)	0.233 (0.06)	**0.029 (0.01)**	**0.016 (0.00)**
CQC simultaneous screen	0.248 (0.07)	0.030 (0.01)	0.017 (0.00)
CQC simultaneous augmented (0.01)	0.245 (0.07)	0.030 (0.01)	0.018 (0.01)
CQC simultaneous augmented (0.05)	0.237 (0.05)	**0.029 (0.01)**	**0.016 (0.00)**
CQC simultaneous screen augmented	0.244 (0.07)	**0.029 (0.01)**	0.017 (0.00)

APPENDIX B: DAY-SPECIFIC PROBABILITIES OF CONCEPTION TECHNICAL DETAILS

B.1 Complete derivations from the main paper

In this section we provide some of the details omitted from the main paper.

B.1.1 Updating step for the fertile window day coefficients and prior parameters

B.1.1.1 Updating step for γ_k

We use a Metropolis step to update γ_k for $k \neq m$. We have the logarithm of the acceptance ratio r given by

$$\begin{aligned}
\log r &= \log \frac{\pi\Big(\gamma_k^* \mid \boldsymbol{Y}, \boldsymbol{W}, \boldsymbol{\gamma}_{-k}, \boldsymbol{\xi}, \boldsymbol{\phi}, m, \mu, \nu, \text{data}\Big)}{\pi\Big(\gamma_k^{(s)} \mid \boldsymbol{Y}, \boldsymbol{W}, \boldsymbol{\gamma}_{-k}, \boldsymbol{\xi}, \boldsymbol{\phi}, m, \mu, \nu, \text{data}\Big)} \\
&= \log \frac{\mathbb{P}\Big(\boldsymbol{W} \mid \gamma_k^*, \boldsymbol{\gamma}_{-k}, \boldsymbol{\xi}, \text{data}\Big)}{\mathbb{P}\Big(\boldsymbol{W} \mid \gamma_k^{(s)}, \boldsymbol{\gamma}_{-k}, \boldsymbol{\xi}, \text{data}\Big)} \frac{\pi\Big(\gamma_k^* \mid m, \mu, \nu\Big)}{\pi\Big(\gamma_k^{(s)} \mid m, \mu, \nu\Big)}.
\end{aligned}$$

For any i, j, k such that $X_{ijk} = 1$ we have that

$$\begin{aligned}
&\log \mathbb{P}\Big(W_{ijk} \mid \gamma_k^*, \boldsymbol{\gamma}_{-k}, \xi_i, \text{data}\Big) \\
&\quad = \log \left\{ \frac{1}{W_{ijk}!} \left[\xi_i \exp\left(\log \gamma_k^* + \sum_{\ell \neq k} u_{ijk\ell} \log \gamma_\ell \right) \right]^{W_{ijk}} \right. \\
&\qquad\qquad \left. \times \ \exp\left\{ -\xi_i \exp\left(\log \gamma_k^* + \sum_{\ell \neq k} u_{ijk\ell} \log \gamma_\ell \right) \right\} \right\} \\
&\quad = -\log(W_{ijk}!) + W_{ijk} \log \xi_i + W_{ijk} \log \gamma_k^* + W_{ijk} \sum_{\ell \neq k} u_{ijk\ell} \log \gamma_\ell \\
&\qquad - \xi_i \exp\left(\log \gamma_k^* + \sum_{\ell \neq k} u_{ijk\ell} \log \gamma_\ell \right),
\end{aligned}$$

and similarly for $\log \mathbb{P}\left(W_{ijk} \mid \gamma_k^{(s)}, \boldsymbol{\gamma}_{-k}, \xi_i, \text{data}\right)$. Thus, it follows that

$$\begin{aligned}
&\log \frac{\mathbb{P}\left(\boldsymbol{W} \,\middle|\, \gamma_k^{*}, \boldsymbol{\gamma}_{-k}, \boldsymbol{\xi}, \text{data}\right)}{\mathbb{P}\left(\boldsymbol{W} \,\middle|\, \gamma_k^{(s)}, \boldsymbol{\gamma}_{-k}, \boldsymbol{\xi}, \text{data}\right)} \\
&\quad = \log \frac{\Pi_{i,j\colon X_{ijk}=1}\, \mathbb{P}\left(W_{ijk} \mid \gamma_k^{*}, \boldsymbol{\gamma}_{-k}, \xi_i, \text{data}\right)}{\Pi_{i,j\colon X_{ijk}=1}\, \mathbb{P}\left(W_{ijk} \mid \gamma_k^{(s)}, \boldsymbol{\gamma}_{-k}, \xi_i, \text{data}\right)} \\
&\quad = \sum_{\substack{i,j\colon \\ X_{ijk}=1}} \log \mathbb{P}\left(W_{ijk} \mid \gamma_k^{*}, \boldsymbol{\gamma}_{-k}, \xi_i, \text{data}\right) - \sum_{\substack{i,j\colon \\ X_{ijk}=1}} \log \mathbb{P}\left(W_{ijk} \mid \gamma_k^{(s)}, \boldsymbol{\gamma}_{-k}, \xi_i, \text{data}\right) \\
&\quad = \sum_{\substack{i,j\colon \\ X_{ijk}=1}} \left\{ W_{ijk}\left(\log \gamma_k^{*} - \log \gamma_k^{(s)}\right) \right. \qquad \text{(B.1)} \\
&\qquad\qquad - \xi_i \exp\left(\log \gamma_k^{*} + \sum_{\ell \neq k} u_{ijk\ell} \log \gamma_\ell\right) \\
&\qquad\qquad \left. + \xi_i \exp\left(\log \gamma_k^{(s)} + \sum_{\ell \neq k} u_{ijk\ell} \log \gamma_\ell\right) \right\}.
\end{aligned}$$

Furthermore,

$$\begin{aligned}
\frac{\pi(\gamma_k^{*})}{\pi(\gamma_k^{(s)})} &= \log \frac{\frac{\left(\frac{\nu}{S(k-m)\,\mu}\right)^{\nu}}{\Gamma(\nu)} \left(\gamma_k^{*}\right)^{\nu-1} \exp\left\{-\frac{\nu}{S(k-m)\,\mu}\gamma_k^{*}\right\}}{\frac{\left(\frac{\nu}{S(k-m)\,\mu}\right)^{\nu}}{\Gamma(\nu)} \left(\gamma_k^{(s)}\right)^{\nu-1} \exp\left\{-\frac{\nu}{S(k-m)\,\mu}\gamma_k^{(s)}\right\}} \\
&= (\nu - 1)\left(\log \gamma_k^{*} - \log \gamma_k^{(s)}\right) - \frac{\nu}{S(k-m)\,\mu}\left(\gamma_k^{*} - \gamma_k^{(s)}\right). \qquad \text{(B.2)}
\end{aligned}$$

Thus, we can obtain the value of $\log r$ for γ_k for $k \neq m$ by summing (B.1) and (B.2).

B.1.2 Prior specification of the fertile window peak intensity coefficient

Let us denote the value of the coefficient corresponding to the peak intensity day in the fertile window as μ. In other words, if k indexes the day with the peak intensity then we have $\gamma_k = \mu$. Then we propose using a gamma distribution to model the peak intensity coefficient μ of the form

$$\mu \sim \text{Gamma}\left(a_\mu,\, b_\mu\right).$$

When we are solely modeling the fertile window day coefficients without any other covariates in the model, then we can utilize the historical data from the literature to provide an informative prior for μ in a manner similar to what we did for the fertile window decay function S. However, when are including other covariates in the model then it is difficult to predict how these other covariates will affect the values of the fertile window day coefficients, so in this case we suggest using an noninformative prior.

B.1.2.1 Updating step for μ

We use a Metropolis step to update μ. We have the logarithm of the acceptance ratio r given by

$$\begin{aligned}\log r &= \log \frac{\pi\Big(\mu^* \;\Big|\; \boldsymbol{Y}, \boldsymbol{W}, \boldsymbol{\gamma}_{-m}, \boldsymbol{\xi}, \phi, m\,\nu\Big)}{\pi\Big(\mu^{(s)} \;\Big|\; \boldsymbol{Y}, \boldsymbol{W}, \boldsymbol{\gamma}_{-m}, \boldsymbol{\xi}, \phi, m\,\nu\Big)} \\ &= \log \left\{ \left(\prod_{k \neq m} \frac{\pi\Big(\gamma_k \,|\, m, \mu^*, \nu\Big)}{\pi(\gamma_k \,|\, m, \mu^{(s)}, \nu)} \right) \frac{\pi(\mu^*)}{\pi(\mu^{(s)})} \right\}.\end{aligned}$$

Furthermore we have for $k \neq m$

$$\begin{aligned}&\log \frac{\pi\Big(\gamma_k \,|\, m, \mu^*, \nu\Big)}{\pi\Big(\gamma_k \,|\, m, \mu^{(s)}, \nu\Big)} \\ &\quad = -\nu \left(\log \gamma_k^* - \log \gamma_k^{(s)} \right) - \frac{\nu \gamma_k}{S(d)} \left(\frac{1}{\mu^*} - \frac{1}{\mu^{(s)}} \right),\end{aligned} \tag{B.3}$$

and

$$\begin{aligned}&\log \frac{\pi\left(\mu^* \,|\, a_\mu, b_\mu\right)}{\pi\left(\mu^{(s)} \,|\, a_\mu, b_\mu\right)} \\ &\quad = \log \frac{\frac{b_\mu^{a_\mu}}{\Gamma(a_\mu)} (\mu^*)^{a_\mu - 1} \exp\left\{-b_\mu \mu^*\right\}}{\frac{b_\mu^{a_\mu}}{\Gamma(a_\mu)} (\mu^{(s)})^{a_\mu - 1} \exp\left\{-b_\mu \mu^{(s)}\right\}} \\ &\quad = (a_\mu - 1) \left(\log \mu^* - \log \mu^{(s)} - b_\mu (\mu^* - \mu^{(s)}) \right).\end{aligned} \tag{B.4}$$

Thus we can calculate the logarithm of the acceptance ratio by summing the expression in B.3 for each $k \neq m$ in addition to the results of B.4.

B.1.3 Prior specification of the variance parameter ν

We assume a prior distribution for ν of the form

$$\nu \sim \text{Gamma}(a_\nu,\, b_\nu). \tag{B.5}$$

Since we don't have any sense of what ν should be from study to study we simply choose an uninformative prior.

B.1.3.1 Updating step for ν

We use a Metropolis step to update ν. The logarithm of the acceptance ratio r is given by

$$\begin{aligned}\log r &= \log \frac{\pi\Big(\nu^* \;\Big|\; \boldsymbol{Y}, \boldsymbol{W}, \boldsymbol{\gamma}, \boldsymbol{\xi}, \boldsymbol{\phi}, m, \mu_0, \text{data}\Big)}{\pi\Big(\nu^{(s)} \;\Big|\; \boldsymbol{Y}, \boldsymbol{W}, \boldsymbol{\gamma}, \boldsymbol{\xi}, \boldsymbol{\phi}, m, \mu_0, \text{data}\Big)} \\ &= \log \frac{\Big(\prod_{\ell \neq m} \pi\big(\gamma_\ell \,|\, m, \mu, \nu^*\big)\Big)\, \pi(\nu^*)}{\Big(\prod_{\ell \neq m} \pi\big(\gamma_\ell \,|\, m, \mu, \nu^{(s)}\big)\Big)\, \pi\big(\nu^{(s)}\big)}.\end{aligned}$$

We have

$$\log \frac{\pi\left(\gamma_k \,|\, m, \mu, \nu^*\right)}{\pi\left(\gamma_k \,|\, m, \mu, \nu^{(s)}\right)} \tag{B.6}$$

$$= \log \frac{\frac{\left(\frac{\nu^*}{S(k-m)\,\mu}\right)^{\nu^*}}{\Gamma(\nu^*)} \left(\gamma_k\right)^{\nu^*-1} \exp\left\{-\frac{\nu^*}{S(k-m)\,\mu}\gamma_k\right\}}{\frac{\left(\frac{\nu^{(s)}}{S(k-m)\,\mu}\right)^{\nu^{(s)}}}{\Gamma(\nu^{(s)})} \left(\gamma_k\right)^{\nu^{(s)}-1} \exp\left\{-\frac{\nu^{(s)}}{S(k-m)\,\mu}\gamma_k\right\}} \tag{B.7}$$

$$\begin{aligned}&= \nu^* \log \frac{\nu^*}{S(k-m)\,\mu} - \nu^{(s)} \log \frac{\nu^{(s)}}{S(k-m)\,\mu} \\ &\qquad - \log\Gamma(\nu^*) + \log\Gamma(\nu^{(s)}) + (\nu^* - \nu^{(s)}) \log \gamma_k \\ &\qquad - (\nu^* - \nu^{(s)}) \frac{\gamma_k}{S(k-m)\,\mu}\end{aligned} \tag{B.8}$$

$$\begin{aligned}&= \nu^* \log \nu^* - \nu^{(s)} \log \nu^{(s)} - (\nu^* - \nu^{(s)}) \log(S(k-m)\,\mu) \\ &\qquad - \log\Gamma(\nu^*) + \log\Gamma(\nu^{(s)}) + (\nu^* - \nu^{(s)}) \log \gamma_k \\ &\qquad - (\nu^* - \nu^{(s)}) \frac{\gamma_k}{S(k-m)\,\mu}\end{aligned} \tag{B.9}$$

$$= \nu^* \log \nu^* - \nu^{(s)} \log \nu^{(s)} - \log \Gamma(\nu^*) + \log \Gamma(\nu^{(s)}) \tag{B.10}$$

$$(\nu^* - \nu^{(s)}) \left(\log \gamma_k - \log \big(S(k-m)\, \mu \big) - \frac{\gamma_k}{S(k-m)\, \mu} \right),$$

and

$$\log \frac{\pi\left(\nu^* \mid a_\nu, b_\nu\right)}{\pi\left(\nu^{(s)} \mid a_\nu, b_\nu\right)}$$

$$= \log \frac{\frac{b_\nu^{a_\nu}}{\Gamma(a_\nu)} (\nu^*)^{a_\nu - 1} \exp\left\{-b_\nu \nu^*\right\}}{\frac{b_\nu^{a_\nu}}{\Gamma(a_\nu)} (\nu^{(s)})^{a_\nu - 1} \exp\left\{-b_\nu \nu^{(s)}\right\}}$$

$$= (a_\nu - 1) \Big(\log \nu^* - \log \nu^{(s)} - b_\nu (\nu^* - \nu^{(s)}) \Big). \tag{B.11}$$

Thus, we can obtain the value of $\log r$ for γ_ℓ by summing (B.10) for $k \neq m$ with and (B.11).

B.1.4 Prior specification of the index of the peak intensity day

We propose modeling the peak intensity day using a multinomial distribution, with fixed prior probabilities chosen by the researcher based on their level of confidence in the scheme used to identify the ovulation day for the data in hand. For example, when ovulation day is determined using a relatively accurate measure such as an ovulation predictor kit in a preponderance of the data, then a researcher may want to place the bulk of the mass in the three or four days around the ovulation day. One the other hand, when the ovulation day is primarily identified by less accurate methods such as by counting backward from the next occurrence of menstrual bleeding, then a researcher may wish to place relatively flat prior probabilities on the distribution.

B.1.4.1 Updating the peak intensity day index

The full conditional distribution for peak intensity day index is given by

$$\mathbb{P}\Big(m = m^* \;\big|\; \boldsymbol{Y}, \boldsymbol{W}, \boldsymbol{\gamma}, \boldsymbol{\xi}, \boldsymbol{\phi}, \mu, \nu, \text{data} \Big)$$

$$= \frac{\Big(\prod_{k \neq m^*} \pi\big(\gamma_k \,|\, m = m^* \mu, \nu\big) \Big)\, \mathbb{P}\left(m = m^*\right)}{\sum_{\ell=1}^{K} \Big(\prod_{k \neq \ell} \pi\big(\gamma_k \,|\, m = \ell, \mu, \nu\big) \Big)\, \mathbb{P}\left(m = \ell\right)},$$

where

$$\pi\big(\gamma_k \,|\, m = \ell, \mu, \nu\big) = \frac{\left(\frac{\nu}{S(k-\ell)\,\mu}\right)^{\nu}}{\Gamma(\nu)} \left(\gamma_k\right)^{\nu-1} \exp\left\{-\frac{\nu}{S(k-\ell)\,\mu}\gamma_k\right\},$$

and

$$\mathbb{P}\left(m = \ell\right) = p_\ell.$$

B.1.5 Gibbs step for missing covariates

Missingness in the covariates can come in the form of a day-specific variable, a cycle-specific variable, or a baseline variable. Let h index the h^{th} element in the ijk^{th} covariate vector, and consider for a moment a baseline variable for participant i. Then there are a set of entries $\{U_{ijkh}\}$ in the design matrix that correspond to the value of this variable. Let us denote such a baseline variable as $U_{i\cdot\cdot}$, then in the case of a continuous variable we have $U_{ijkh} = U_{i\cdot\cdot}$ for the corresponding set of j, k, and h. When the variable is a categorical variable, we have design matrix dummy variables $(U_{ijk\ell_1}, \ldots U_{ijk\ell_r})$ for some $\ell_1, \ldots, \ell_r$ for each cycle j and fertile window day k corresponding to the value of $U_{i\cdot\cdot}$ for participant i. We similarly denote $U_{ij\cdot}$ to be a cycle-specific variable with corresponding entries in the design matrix $\{U_{ijkh}\}$ for the set of k and h.

We assume that the prior distributions for the covariates are mutually independent. Although we acknowledge that this is not a realistic assumption for every setting, our contention is that without study-specific knowledge of the variables involved, it is difficult to specify a reasonable joint distribution model. We expect this to be a reasonable approximation for what is ideally a relatively small amount of missingness in the data.

The model for missing continuous covariates and missing categorical covariates differ slightly, as we will cover in the following sections.

B.1.5.1 Gibbs step for missing continuous covariate

Let U_{ijkh} be a missing day-specific covariate, then we use a Metropolis step to update U_{ijkh}. Under the previously stated assumptions, we have

$$\pi\Big(U_{ijkh} \;\Big|\; \boldsymbol{Y}, \boldsymbol{W}, \boldsymbol{\gamma}, \boldsymbol{\xi}, \phi, \boldsymbol{U}_{-ijkh}\Big)$$

$$= \frac{\pi\Big(\boldsymbol{Y}, \boldsymbol{W}, \boldsymbol{\gamma}, \boldsymbol{\xi}, \phi, U_{ijkh}, \boldsymbol{U}_{-ijkh}\Big)}{\int \pi\Big(\boldsymbol{Y}, \boldsymbol{W}, \boldsymbol{\gamma}, \boldsymbol{\xi}, \phi, U_{ijkh}, \boldsymbol{U}_{-ijkh}\Big)\, dU_{ijkh}}$$

$$\propto \frac{\pi\Big(\boldsymbol{W}_i \,\Big|\, \boldsymbol{\gamma}, \xi_i, \boldsymbol{X}, U_{ijkh}, \boldsymbol{U}_{-ijkh}\Big)\, \mathbb{P}\left(X_{ijk} \,|\, U_{ijkh}\right)\, \pi(U_{ijkh})}{\int \pi\Big(\boldsymbol{W}_i \,\Big|\, \boldsymbol{\gamma}, \xi_i, \boldsymbol{X}, U_{ijkh}, \boldsymbol{U}_{-ijkh}\Big)\, \mathbb{P}\left(X_{ijk} \,|\, U_{ijkh}\right)\, \pi(U_{ijkh})\, dU_{ijkh}},$$

where the $\mathbb{P}\left(\boldsymbol{X}_{ijk} \,|\, U_{ijk}\right)$ term is a constant if X_{ijk} is nonmissing for the i^{th} subject. It follows that the acceptance ratio r is given by

$$r_{\text{day-specific}}$$

$$= \frac{\pi\Big(U_{ijkh} = u^{*} \,\Big|\, \boldsymbol{Y}, \boldsymbol{W}, \boldsymbol{\gamma}, \boldsymbol{\xi}, \phi, \boldsymbol{U}_{-ijkh}\Big)}{\pi\Big(U_{ijkh} = u^{(s)} \,\Big|\, \boldsymbol{Y}, \boldsymbol{W}, \boldsymbol{\gamma}, \boldsymbol{\xi}, \phi, \boldsymbol{U}_{-ijkh}\Big)}$$

$$= \frac{\pi\Big(\boldsymbol{W}_i \,\Big|\, \boldsymbol{\gamma}, \xi_i, \boldsymbol{X}, U_{ijkh} = u^{*}, \boldsymbol{U}_{-ijkh}\Big)\, \mathbb{P}\left(X_{ijk} \,|\, U_{ijkh} = u^{*}\right)\, \pi(U_{ijkh} = u^{*})}{\pi\Big(\boldsymbol{W}_i \,\Big|\, \boldsymbol{\gamma}, \xi_i, \boldsymbol{X}, U_{ijkh} = u^{(s)}, \boldsymbol{U}_{-ijkh}\Big)\, \mathbb{P}\left(X_{ijk} \,|\, U_{ijkh} = u^{(s)}\right)\, \pi(U_{ijkh} = u^{(s)})}.$$

Next, suppose that the missing covariate is a cycle-specific covariate, then let us denote the random variable corresponding to this value as $U_{ij\cdot h}$. Then $U_{ijkh} = U_{ij\cdot h}$ for all fertile window days $k = 1, \dots, K$, and the acceptance ratio is given by

$$r_{\text{cycle-specific}}$$

$$= \frac{\pi\Big(U_{ij\cdot h} = u^{*} \,\Big|\, \boldsymbol{Y}, \boldsymbol{W}, \boldsymbol{\gamma}, \boldsymbol{\xi}, \phi, \boldsymbol{U}_{-ij\cdot h}\Big)}{\pi\Big(U_{ij\cdot h} = u^{(s)} \,\Big|\, \boldsymbol{Y}, \boldsymbol{W}, \boldsymbol{\gamma}, \boldsymbol{\xi}, \phi, \boldsymbol{U}_{-ij\cdot h}\Big)}$$

$$= \frac{\pi\Big(\boldsymbol{W}_{ij} \,\Big|\, \boldsymbol{\gamma}, \xi_i, \boldsymbol{X}, U_{ij\cdot h} = u^{*}, \boldsymbol{U}_{-ij\cdot h}\Big)\, \mathbb{P}\left(\boldsymbol{X}_{ij} \,|\, U_{ij\cdot h} = u^{*}\right)\, \pi(U_{ij\cdot h} = u^{*})}{\pi\Big(\boldsymbol{W}_{ij} \,\Big|\, \boldsymbol{\gamma}, \xi_i, \boldsymbol{X}, U_{ij\cdot h} = u^{(s)}, \boldsymbol{U}_{-ij\cdot h}\Big)\, \mathbb{P}\left(\boldsymbol{X}_{ij} \,|\, U_{ij\cdot h} = u^{(s)}\right)\, \pi(U_{ij\cdot h} = u^{(s)})},$$

Finally, suppose that the missing covariate is a baseline covariate, then let us denote the random variable corresponding to this value as $U_{i\cdot\cdot h}$. Then $U_{ijkh} = U_{i\cdot\cdot h}$ for all cycles j and fertile window days k associated with participant i, and the acceptance ratio is given by

$$r_{\text{baseline}}$$

$$= \frac{\pi\Big(U_{i\cdot\cdot h} = u^* \,\Big|\, \boldsymbol{Y}, \boldsymbol{W}, \boldsymbol{\gamma}, \boldsymbol{\xi}, \phi, \boldsymbol{U}_{-i\cdot\cdot h}\Big)}{\pi\Big(U_{i\cdot\cdot h} = u^{(s)} \,\Big|\, \boldsymbol{Y}, \boldsymbol{W}, \boldsymbol{\gamma}, \boldsymbol{\xi}, \phi, \boldsymbol{U}_{-i\cdot\cdot h}\Big)}$$

$$= \frac{\pi\Big(\boldsymbol{W}_i \,\Big|\, \boldsymbol{\gamma}, \xi_i, \boldsymbol{X}, U_{i\cdot\cdot h} = u^*, \boldsymbol{U}_{-i\cdot\cdot h}\Big)\, \mathbb{P}\left(\boldsymbol{X}_i \,|\, U_{i\cdot\cdot h} = u^*\right)\, \pi(U_{i\cdot\cdot h} = u^*)}{\pi\Big(\boldsymbol{W}_i \,\Big|\, \boldsymbol{\gamma}, \xi_i, \boldsymbol{X}, U_{i\cdot\cdot h} = u^{(s)}, \boldsymbol{U}_{-i\cdot\cdot h}\Big)\, \mathbb{P}\left(\boldsymbol{X}_i \,|\, U_{i\cdot\cdot h} = u^{(s)}\right)\, \pi(U_{i\cdot\cdot h} = u^{(s)}).}.$$

Additionally, we also need to specify a prior distribution for the covariates. For this paper we used a conjugate prior Normal distribution for the imputed covariate. In more detail, we specify that

$$\begin{aligned} U_{ijkh} \,|\, \zeta, \kappa &\overset{i.i.d.}{\sim} N(\zeta, \kappa), \qquad \text{for all } i, j, k, h, \\ \zeta \,|\, \kappa, \{U_{ijkh}\} &\sim N(\zeta_0, \kappa), \\ \kappa \,|\, \{U_{ijkh}\} &\sim \text{Gamma}(a_\kappa, b_\kappa), \end{aligned}$$

where $\{U_{ijkh}\}$ is taken to mean the set of all missing h^{th} covariates. We take an Empirical Bayes approach when specifying the hyperparameters. We set ζ_0 to be the empirical mean of the nonmissing h^{th} covariates, and select a_κ, and b_κ so that the prior distribution's mean is equal to the sample variance of the nonmissing h^{th} covariates and let the distribution have a noninformative variance. We use a uniform distribution as the proposal distribution for the missing covariates.

B.1.5.2 Gibbs step for missing categorical covariate

Let U_{ijk} be a day-specific categorical variable with corresponding design matrix dummy variables $(U_{ijk\ell_1}, \ldots U_{ijk\ell_r})$. Then

$$\begin{aligned} &\mathbb{P}\left(U_{ijk} = u \,\Big|\, \boldsymbol{Y}, \boldsymbol{W}, \boldsymbol{\gamma}, \boldsymbol{\xi}, \phi, \boldsymbol{U}_{-ijk}\right) \\ &\quad = \frac{\pi\Big(\boldsymbol{Y}, \boldsymbol{W}, \boldsymbol{\gamma}, \boldsymbol{\xi}, \phi, U_{ijk} = u, \boldsymbol{U}_{-ijk}\Big)}{\sum_t \pi\Big(\boldsymbol{Y}, \boldsymbol{W}, \boldsymbol{\gamma}, \boldsymbol{\xi}, \phi, U_{ijk} = t, \boldsymbol{U}_{-ijk}\Big)} \\ &\quad = \frac{\mathbb{P}\left(W_{ijk} \,\Big|\, \boldsymbol{\gamma}, \xi_i, \boldsymbol{X}, U_{ijk} = u, \boldsymbol{U}_{-ijk}\right) \mathbb{P}\left(X_{ijk} \,|\, X_{ij,k-1}, U_{ijk} = u\right) \mathbb{P}\left(U_{ijk} = u\right)}{\sum_t \mathbb{P}\left(W_{ijk} \,\Big|\, \boldsymbol{\gamma}, \xi_i, \boldsymbol{X}, U_{ijk} = t, \boldsymbol{U}_{-ijk}\right) \mathbb{P}\left(X_{ijk} \,|\, X_{ij,k-1}, U_{ijk} = t\right) \mathbb{P}\left(U_{ijk} = t\right)}. \end{aligned}$$

Next, let $U_{ij\cdot}$ be a cycle-specific categorical variable with corresponding design matrix dummy variables $\{(U_{ijk\ell_1},\ldots,U_{ijk\ell_r})\}$. Then

$$\begin{aligned}
&\mathbb{P}\left(U_{ij\cdot}=u \,\middle|\, \boldsymbol{Y},\boldsymbol{W},\boldsymbol{\gamma},\boldsymbol{\xi},\phi,\boldsymbol{U}_{-ij\cdot}\right)\\
&\qquad=\frac{\pi\Big(\boldsymbol{Y},\boldsymbol{W},\boldsymbol{\gamma},\boldsymbol{\xi},\phi,U_{ij\cdot}=u,\boldsymbol{U}_{-ij\cdot}\Big)}{\sum_t\pi\Big(\boldsymbol{Y},\boldsymbol{W},\boldsymbol{\gamma},\boldsymbol{\xi},\phi,U_{ij\cdot}=t,\boldsymbol{U}_{-ij\cdot}\Big)}\\
&\qquad=\frac{\mathbb{P}\left(\boldsymbol{W}_{ij}\,\middle|\,\boldsymbol{\gamma},\xi_i,\boldsymbol{X},U_{ij\cdot}=u,\boldsymbol{U}_{-ij\cdot}\right)\mathbb{P}\left(\boldsymbol{X}_{ij}\,|\,U_{ij\cdot}=u\right)\mathbb{P}\left(U_{ij\cdot}=u\right)}{\sum_t\mathbb{P}\left(\boldsymbol{W}_{ij}\,\middle|\,\boldsymbol{\gamma},\xi_i,\boldsymbol{X},U_{ij\cdot}=t,\boldsymbol{U}_{-ij\cdot}\right)\mathbb{P}\left(\boldsymbol{X}_{ij}\,|\,U_{ij\cdot}=t\right)\mathbb{P}\left(U_{ij\cdot}=t\right)}.
\end{aligned}$$

Finally, let $U_{i\cdot\cdot}$ be a baseline categorical variable with corresponding design matrix dummy variables $\{(U_{ijk\ell_1},\ldots,U_{ijk\ell_r})\}$. Then

$$\begin{aligned}
&\mathbb{P}\left(U_{i\cdot\cdot}=u \,\middle|\, \boldsymbol{Y},\boldsymbol{W},\boldsymbol{\gamma},\boldsymbol{\xi},\phi,\boldsymbol{U}_{-ij\cdot}\right)\\
&\qquad=\frac{\pi\Big(\boldsymbol{Y},\boldsymbol{W},\boldsymbol{\gamma},\boldsymbol{\xi},\phi,U_{i\cdot\cdot}=u,\boldsymbol{U}_{-ij\cdot}\Big)}{\sum_t\pi\Big(\boldsymbol{Y},\boldsymbol{W},\boldsymbol{\gamma},\boldsymbol{\xi},\phi,U_{i\cdot\cdot}=t,\boldsymbol{U}_{-ij\cdot}\Big)}\\
&\qquad=\frac{\mathbb{P}\left(\boldsymbol{W}_{i}\,\middle|\,\boldsymbol{\gamma},\xi_i,\boldsymbol{X},U_{i\cdot\cdot}=u,\boldsymbol{U}_{-ij\cdot}\right)\mathbb{P}\left(\boldsymbol{X}_{i}\,|\,U_{i\cdot\cdot}=u\right)\mathbb{P}\left(U_{i\cdot\cdot}=u\right)}{\sum_t\mathbb{P}\left(\boldsymbol{W}_{i}\,\middle|\,\boldsymbol{\gamma},\xi_i,\boldsymbol{X},U_{i\cdot\cdot}=t,\boldsymbol{U}_{-ij\cdot}\right)\mathbb{P}\left(\boldsymbol{X}_{i}\,|\,U_{i\cdot\cdot}=t\right)\mathbb{P}\left(U_{i\cdot\cdot}=t\right)}.
\end{aligned}$$

We specify the prior distribution to be conjugate multinomial-Dirichlet distributed. That is to say we let

$$\begin{aligned}
U_{ijkh}\,|\,\boldsymbol{\pi}_h &\sim \text{multinomial}(\boldsymbol{\pi}_h),\\
\boldsymbol{\pi}_h &\sim \text{Dirichlet}(\boldsymbol{a}_h).
\end{aligned}$$

We use an empirical Bayes approach to specify the mean of the prior distributions of the $\boldsymbol{a}_h$.

B.2 Detailed simulation studies

B.2.1 Simulation setup and model comparisons

In order to compare the peak-intensity decay model, we compare the peak-intensity decay model with two versions of the day-specific probabilities model proposed in Dunson and Stanford (2005). In the first version of the day-specific probabilities model, we use the data that we generated in Section 3 of the main paper to provide informative priors to the model coefficients. Then, for the second specification of the model we used uninformative priors.

In order to compare these various models, we considered a number of different settings. In terms of the simulated data settings, we consider fertile window sizes of both five days and seven days. In terms of the sample size we considered sizes of 100, 200, and 300 participants. For the coefficients corresponding to the fertile window days, we consider three settings.

In the first setting, we set the coefficients to be the same as the decay that we derived in Section 3 of the main paper, so that both the peak-intensity decay model and the informative priors day-specific probabilites models are correctly specified. In another setting, we set the peak-intensity of the coefficients corresponding to the fertile window to be scaled from that of the prior distributions, and we shifted the location of the peak-intensity day. We would expect that the peak-intensity decay model would be able to recover these parameters more readily than the incorrectly specified informative priors day-specific probabilities model. Finally, we also included a model in which the decay of the coefficients corresponding to the fertile window days is different than what was derived in 3 of the main paper, so that both the peak-intensity decay model and the informative priors day-specific probabilities model are incorrectly specified.

We present the results from a pair of simulation studies in what follows, and in the Supplementary Materials we present an extended set of simulation results. For each of the simulation studies we are describing the average results over 10,000 simulations.

B.2.2 Correctly specified prior parameters setting

In Tables B.1, B.2, and B.3, we consider the setting where the data were generated based upon coefficients constructed from the elicited decay curve created in Section 3 of the main paper with a 5-day fertile window, and in Tables B.10, B.11, and B.11 we consider the same setting but with a 7-day fertile window.

In Table B.1, we compare the models for a 5-day fertile window with 100 participants. With this small number of participants, the peak-intensity model has a large amount of uncertainty regarding the location of the peak intensity location, with a 0.172 posterior probability of Day -1 being the peak intensity day, and a 0.351 posterior probability of Day 1 being the peak intensity day. As a result, the posterior mean coefficient value of the peak intensity day is slightly lower than the true value (-0.025 bias). However, despite this small bias and uncertainty in the peak intensity day location, the MSEs for the peak intensity model coefficients are still smaller than for the other models, so we conclude that the posterior distribution for the peak-intensity model has a higher posterior density near the true values than for the other models.

In Table B.10, we again see that there is a large amount of uncertainty regarding the location of the peak intensity location (a 0.209 posterior probability of Day -1 being the peak intensity day, and a 0.351 posterior probability of Day 1 being the peak intensity day). The bias for the coefficient posterior mean for the peak intensity day is -0.051 in this setting, but again we see that the MSEs for the peak intensity model coefficients are smaller than for the other models.

In Tables B.2 and B.11, we compare the models for a 5-day and 7-day fertile window with 200 participants. For the 5-day version, the posterior probability for the peak intensity day is reduced to 0.080 for Day -1 and to 0.287 for Day 1, and the bias for the coefficient posterior mean for the peak intensity day is improved to -0.009. The width of the credible region for the peak-intensity model for each of the coefficients is smaller than for either of the other models. For example, the width of the credible region for the coefficient posterior mean for the peak intensity day is 0.218, compared to 0.278 and 0.289 for the fixed informative and non-informative models.

For Table B.11, the posterior probability for the peak intensity day is reduced to 0.114 for Day -1 and to 0.165 for Day 1, and the bias for the coefficient posterior mean for the peak intensity day is improved to -0.018. the width of the credible region for the coefficient posterior mean for the peak intensity day is 0.225, compared to 0.303 and 0.321 for the fixed informative and non-informative models.

In Tables B.3 and B.12, we compare the models for a 5-day and 7-day fertile window with 300 participants. With this many participants the posterior probability for the peak intensity day location is 0.930 for the 5-day fertile window setting, and 0.844 for the 7-day fertile window setting. With this many participants, all three of the models have coefficient means that are close to the true values. The credible region widths are still the smallest for the peak-intensity model, but the difference between the widths across the models is slightly smaller. For example, the width of the credible region is 0.91 for the peak intensity coefficient for the peak intensity model, compared to 0.233 for the fixed informative model and 0.240 for the non-informative model. When the study size is smaller, the information borrowing that is enabled by the peak intensity model seems to provide greater relative gains, but even when data are more abundant the models without the information sharing are still less efficient.

B.2.3 Scaled and shifted prior parameters setting

In Tables B.4, B.5, and B.6, we consider the setting where the data were generated based upon coefficients from a scaled and shifted decay curve as compared to the elicited decay curve created in Section 3 of the main paper with a 5-day fertile window, and in Tables B.13, B.14, and B.15 we consider the same setting but with a 7-day fertile window.

In Table B.4 we compare the models for a 5-day and 7-day fertile window with 300 participants. In this setting were we have a relatively small amount of data we would expect that the peak-intensity and the fixed-informative models would be guided to some degree by the misspecified prior parameters, and we see that to some degree, particularly in the negative bias of the posterior mean value for the peak intensity day (-0.035 for the peak-intensity model and -0.043 for the fixed informative model). As we saw in the correctly specified models, there is a fair amount of uncertainty in the posterior distribution of the peak intensity day. Interestingly though, the average posterior proportion for the peak intensity day for this setting with a shifted peak intensity day was 0.569 compared to that of the 0.468 for the correctly specified setting with the same number of participants. We attribute this behavior to the fact that the scaled true coefficient values make it comparatively easier to distinguish the peak coefficient day. The same is true for the 7-day fertile window setting in Table B.13 where the average posterior proportion for the peak intensity day was 0.773, but in this case the cause is due to the longer fertile window in the simulation setting resulting in more pregnancy events.

In Tables B.4 and B.13 we also see that the flexibility of the peak-intensity model in terms of being able to adapt to location and scale shifts results in posterior densities with means closer to the true parameters than for the other models in most cases. For example, the peak intensity model has an average posterior mean bean of -0.012 and credible region width of 0.325 for Day -1 compared to -.050 and -0.019 with credible region widths of 0.379 and 0.422 for the fixed informative and non-informative models. In Tables B.4 and B.13 we see similar results for the average posterior means for the Day -1 coefficient, where the peak intensity model has a bias of -0.012 with a credible region widths of 0.244, while the fixed informative and non-informative models have biases of -0.034 and -0.016 with credible region widths of 0.294 and 0.311.

In Tables B.6 and B.15 with 300 participants, all of the models are performing well. However, the peak intensity model tends to have the smallest posterior variance for the coefficients in comparison to the other models. For example, the credible region width for the peak intensity day coefficient is 0.222 for the peak intensity model compared to 0.269 and 0.278 for the fixed informative and non-informative models.

B.2.4 Incorrect decay prior parameters setting

The incorrect decay simulation setting is a setting that is poorly suited for the peak intensity model, since it violates one of the core assumptions of the model. In particular, the simulated data has a local minimum at Day -1 in both the 5-day and 7-day simulation settings. Since the other models don't rely on any particular form of the fertile window coefficients, they shouldn't suffer as greatly in this situation. Although we include such a setting in our simulation studies so that we can understand the behavior of the peak-intensity model when the true fertile window decay curve doesn't follow the one specified by the prior parameters, it is our belief that since the biology of human fertility and empirical evidence from past studies dictates such a form of the decay curve, that it is reasonable to impose such a form to the model in order to incur gains in efficiency in the event that such beliefs are reasonable close to the truth.

In Table B.7, we see that the peak intensity model has a large amount of uncertainty in the posterior distribution of the peak intensity day location for 100 participants, with proportions of 0.125, 0.368, 0.337, and 0.159 for Day -1, Day 0, Day 1, and Day 2. The average posterior mean bias for the Day -1 coefficient for the peak intensity model is 0.039 compared to 0.008 for both the fixed informative and non-informative models. The 7-day fertile window setting shown in Table B.16, shows similar results, except that the other models exhibit worse performance with more days to estimate. For example, the fixed informative model has a posterior mean bias of -0.031 for Day -1 compared to 0.007 in the 5-day fertile window setting.

In Table B.8 we see the results for 200 participants in the 5-day fertile window setting. With this increase in participants, although the bias is still worse than for the fixed informative and non-informative models for the Day -2 and Day -1 coefficients, the difference between the posterior means for Day -2 and Day -1 is -0.022 compared to -0.033 for the 100 participants setting. Then in Table B.9, the difference between the posterior means for Day -2 and Day -1 is -0.009, so we can see that the model is adjusting to the local minimum at Day -1.

In Tables B.17 and B.18 we see the results for 200 and 300 participants in the 7-day fertile window setting. In both tables we do see that the peak intensity model not well-suited to model the local minimum at Day -1. For example, the difference between the posterior means for Day -2 and Day -1 is -0.032 and -0.021 for the two settings.

Table B.1: Correctly specified prior parameters, 5 fertile window days, $n = 100$

day	true	mean	bias	MSE	Std.	CR
			Peak-intensity			
Day -3	0.115	0.108	-0.007	0.001	0.040	0.156
Day -2	0.196	0.182	-0.014	0.002	0.053	0.208
Day -1	0.288	0.267	-0.021	0.003	0.062	0.241
Day 0	0.340	0.315	-0.025	0.004	0.069	0.271
Day 1	0.264	0.262	-0.002	0.006	0.085	0.329
			Fixed informative			
Day -3	0.115	0.093	-0.022	0.004	0.058	0.213
Day -2	0.196	0.183	-0.013	0.005	0.075	0.287
Day -1	0.288	0.280	-0.008	0.006	0.087	0.339
Day 0	0.340	0.334	-0.006	0.007	0.093	0.364
Day 1	0.264	0.259	-0.005	0.005	0.085	0.329
			Non-informative			
Day -3	0.115	0.127	0.012	0.005	0.071	0.264
Day -2	0.196	0.188	-0.008	0.007	0.082	0.314
Day -1	0.288	0.268	-0.020	0.009	0.094	0.365
Day 0	0.340	0.314	-0.026	0.010	0.100	0.389
Day 1	0.264	0.252	-0.012	0.008	0.092	0.355

Posterior distribution for the peak intensity

true	mean	bias	50% CR	90% CR
0.340	0.349	0.009	0.086	0.165

Posterior distribution for the peak
intensity location

Day -3	Day -2	Day -1	Day 0	Day 1
0.000	0.010	0.172	0.468	0.351

Table B.2: Correctly specified prior parameters, 5 fertile window days, $n = 200$

day	true	mean	bias	MSE	Std.	CR
			Peak-intensity			
Day -3	0.115	0.105	-0.010	0.001	0.031	0.121
Day -2	0.196	0.182	-0.014	0.001	0.042	0.162
Day -1	0.288	0.272	-0.016	0.002	0.050	0.195
Day 0	0.340	0.331	-0.009	0.002	0.056	0.218
Day 1	0.264	0.273	0.009	0.004	0.064	0.251
			Fixed informative			
Day -3	0.115	0.102	-0.013	0.002	0.048	0.177
Day -2	0.196	0.185	-0.011	0.003	0.057	0.222
Day -1	0.288	0.279	-0.009	0.004	0.066	0.258
Day 0	0.340	0.336	-0.004	0.004	0.071	0.278
Day 1	0.264	0.266	0.002	0.004	0.066	0.254
			Non-informative			
Day -3	0.115	0.120	0.005	0.003	0.053	0.198
Day -2	0.196	0.188	-0.008	0.004	0.061	0.234
Day -1	0.288	0.273	-0.015	0.005	0.069	0.269
Day 0	0.340	0.326	-0.014	0.006	0.074	0.289
Day 1	0.264	0.263	-0.001	0.004	0.069	0.266

Posterior distribution for the peak intensity

true	mean	bias	50% CR	90% CR
0.340	0.354	0.014	0.075	0.145

Posterior distribution for the peak intensity location

Day -3	Day -2	Day -1	Day 0	Day 1
0.000	0.000	0.080	0.633	0.287

Table B.3: Correctly specified prior parameters, 5 fertile window days, $n = 300$

day	true	mean	bias	MSE	Std.	CR
Peak-intensity						
Day -3	0.115	0.108	-0.007	0.001	0.028	0.107
Day -2	0.196	0.190	-0.006	0.001	0.037	0.145
Day -1	0.288	0.280	-0.008	0.002	0.044	0.173
Day 0	0.340	0.335	-0.005	0.002	0.049	0.191
Day 1	0.264	0.270	0.006	0.003	0.053	0.209
Fixed informative						
Day -3	0.115	0.104	-0.011	0.002	0.041	0.153
Day -2	0.196	0.196	0.000	0.002	0.050	0.193
Day -1	0.288	0.285	-0.003	0.003	0.056	0.219
Day 0	0.340	0.338	-0.002	0.004	0.060	0.233
Day 1	0.264	0.264	0.000	0.003	0.055	0.214
Non-informative						
Day -3	0.115	0.115	0.000	0.002	0.044	0.164
Day -2	0.196	0.198	0.002	0.003	0.052	0.200
Day -1	0.288	0.281	-0.007	0.004	0.058	0.226
Day 0	0.340	0.331	-0.009	0.004	0.061	0.240
Day 1	0.264	0.261	-0.003	0.003	0.057	0.221

Posterior distribution for the peak intensity

true	mean	bias	50% CR	90% CR
0.340	0.356	0.016	0.068	0.133

Posterior distribution for the peak intensity location

Day -3	Day -2	Day -1	Day 0	Day 1
0.000	0.000	0.027	0.930	0.043

Table B.4: Scaled and shifted prior parameters, 5 fertile window days, $n = 100$

day	true	mean	bias	MSE	Std.	CR
				Peak-intensity		
Day -2	0.235	0.235	0.000	0.006	0.074	0.286
Day -1	0.346	0.334	-0.012	0.006	0.083	0.325
Day 0	0.408	0.373	-0.035	0.004	0.078	0.308
Day 1	0.317	0.278	-0.039	0.007	0.077	0.300
Day 2	0.125	0.127	0.002	0.004	0.067	0.251
				Fixed informative		
Day -2	0.235	0.174	-0.061	0.010	0.081	0.307
Day -1	0.346	0.296	-0.050	0.010	0.098	0.379
Day 0	0.408	0.365	-0.043	0.009	0.105	0.407
Day 1	0.317	0.317	0.000	0.007	0.097	0.377
Day 2	0.125	0.173	0.048	0.006	0.077	0.294
				Non-informative		
Day -2	0.235	0.228	-0.007	0.009	0.095	0.363
Day -1	0.346	0.327	-0.019	0.011	0.109	0.422
Day 0	0.408	0.369	-0.039	0.012	0.114	0.444
Day 1	0.317	0.286	-0.031	0.011	0.104	0.400
Day 2	0.125	0.137	0.012	0.006	0.078	0.293

Posterior distribution for the peak intensity

true	mean	bias	50% CR	90% CR
0.340	0.356	0.016	0.068	0.133

Posterior distribution for the peak
intensity location

Day -2	Day -1	Day 0	Day 1	Day 2
0.001	0.232	0.569	0.173	0.025

Table B.5: Scaled and shifted prior parameters, 5 fertile window days, $n = 200$

day	true	mean	bias	MSE	Std.	CR
			Peak-intensity			
Day -2	0.235	0.229	-0.006	0.002	0.052	0.205
Day -1	0.346	0.334	-0.012	0.003	0.062	0.244
Day 0	0.408	0.393	-0.015	0.003	0.064	0.252
Day 1	0.317	0.301	-0.016	0.003	0.059	0.232
Day 2	0.125	0.123	-0.002	0.002	0.043	0.166
			Fixed informative			
Day -2	0.235	0.201	-0.034	0.005	0.065	0.250
Day -1	0.346	0.312	-0.034	0.007	0.075	0.294
Day 0	0.408	0.387	-0.021	0.006	0.081	0.316
Day 1	0.317	0.321	0.004	0.005	0.075	0.291
Day 2	0.125	0.146	0.021	0.003	0.056	0.214
			Non-informative			
Day -2	0.235	0.230	-0.005	0.004	0.070	0.271
Day -1	0.346	0.330	-0.016	0.007	0.080	0.311
Day 0	0.408	0.391	-0.017	0.007	0.085	0.331
Day 1	0.317	0.306	-0.011	0.006	0.078	0.302
Day 2	0.125	0.122	-0.003	0.003	0.057	0.212

Posterior distribution for the peak intensity

true	mean	bias	50% CR	90% CR
0.408	0.403	-0.005	0.074	0.145

Posterior distribution for the peak intensity location

Day -2	Day -1	Day 0	Day 1	Day 2
0.000	0.094	0.806	0.096	0.004

Table B.6: Scaled and shifted prior parameters, 5 fertile window days, $n = 300$

day	true	mean	bias	MSE	Std.	CR
			Peak-intensity			
Day -2	0.235	0.231	-0.004	0.001	0.044	0.172
Day -1	0.346	0.338	-0.008	0.002	0.053	0.208
Day 0	0.408	0.403	-0.005	0.002	0.057	0.222
Day 1	0.317	0.309	-0.008	0.002	0.050	0.198
Day 2	0.125	0.124	-0.001	0.001	0.034	0.131
			Fixed informative			
Day -2	0.235	0.214	-0.021	0.004	0.057	0.218
Day -1	0.346	0.323	-0.023	0.004	0.064	0.251
Day 0	0.408	0.397	-0.011	0.005	0.069	0.269
Day 1	0.317	0.318	0.001	0.004	0.063	0.246
Day 2	0.125	0.143	0.018	0.002	0.048	0.181
			Non-informative			
Day -2	0.235	0.235	0.000	0.004	0.059	0.229
Day -1	0.346	0.335	-0.011	0.004	0.067	0.261
Day 0	0.408	0.401	-0.007	0.006	0.071	0.278
Day 1	0.317	0.307	-0.010	0.004	0.065	0.253
Day 2	0.125	0.127	0.002	0.002	0.048	0.182

Posterior distribution for the peak intensity

true	mean	bias	50% CR	90% CR
0.408	0.406	-0.002	0.066	0.129

Posterior distribution for the peak intensity location

Day -2	Day -1	Day 0	Day 1	Day 2
0.000	0.035	0.912	0.053	0.001

Table B.7: Incorrect decay prior parameters, 5 fertile window days, $n = 100$

day	true	mean	bias	MSE	Std.	CR
			Peak-intensity			
Day -2	0.180	0.116	-0.064	0.007	0.045	0.172
Day -1	0.110	0.149	0.039	0.003	0.045	0.175
Day 0	0.200	0.193	-0.007	0.001	0.046	0.181
Day 1	0.180	0.177	-0.003	0.002	0.048	0.188
Day 2	0.140	0.121	-0.019	0.004	0.056	0.211
			Fixed informative			
Day -2	0.180	0.133	-0.047	0.006	0.061	0.228
Day -1	0.110	0.118	0.008	0.002	0.056	0.211
Day 0	0.200	0.202	0.002	0.003	0.067	0.260
Day 1	0.180	0.200	0.020	0.003	0.066	0.255
Day 2	0.140	0.155	0.015	0.003	0.061	0.231
			Non-informative			
Day -2	0.180	0.172	-0.008	0.005	0.069	0.264
Day -1	0.110	0.118	0.008	0.004	0.060	0.225
Day 0	0.200	0.188	-0.012	0.005	0.071	0.273
Day 1	0.180	0.169	-0.011	0.004	0.068	0.262
Day 2	0.140	0.138	-0.002	0.004	0.063	0.239

Posterior distribution for the peak intensity

true	mean	bias	50% CR	90% CR
0.200	0.235	0.035	0.060	0.117

Posterior distribution for the peak intensity location

Day -2	Day -1	Day 0	Day 1	Day 2
0.010	0.125	0.368	0.337	0.159

Table B.8: Incorrect decay prior parameters, 5 fertile window days, $n = 200$

day	true	mean	bias	MSE	Std.	CR
			Peak-intensity			
Day -2	0.180	0.122	-0.058	0.005	0.036	0.139
Day -1	0.110	0.144	0.034	0.002	0.035	0.137
Day 0	0.200	0.204	0.004	0.001	0.039	0.152
Day 1	0.180	0.190	0.010	0.001	0.039	0.152
Day 2	0.140	0.125	-0.015	0.003	0.044	0.165
			Fixed informative			
Day -2	0.180	0.154	-0.026	0.003	0.048	0.184
Day -1	0.110	0.113	0.003	0.001	0.042	0.161
Day 0	0.200	0.201	0.001	0.002	0.051	0.198
Day 1	0.180	0.193	0.013	0.002	0.050	0.193
Day 2	0.140	0.149	0.009	0.002	0.046	0.176
			Non-informative			
Day -2	0.180	0.174	-0.006	0.003	0.052	0.197
Day -1	0.110	0.113	0.003	0.002	0.044	0.168
Day 0	0.200	0.195	-0.005	0.003	0.053	0.205
Day 1	0.180	0.177	-0.003	0.003	0.051	0.198
Day 2	0.140	0.140	0.000	0.002	0.048	0.182

Posterior distribution for the peak intensity

true	mean	bias	50% CR	90% CR
0.200	0.238	0.038	0.049	0.097

Posterior distribution for the peak intensity location

Day -2	Day -1	Day 0	Day 1	Day 2
0.000	0.032	0.487	0.409	0.072

Table B.9: Incorrect decay prior parameters, 5 fertile window days, $n = 300$

day	true	mean	bias	MSE	Std.	CR
			Peak-intensity			
Day -2	0.180	0.130	-0.050	0.004	0.033	0.127
Day -1	0.110	0.139	0.029	0.002	0.031	0.120
Day 0	0.200	0.206	0.006	0.001	0.035	0.136
Day 1	0.180	0.189	0.009	0.001	0.034	0.132
Day 2	0.140	0.123	-0.017	0.002	0.037	0.140
			Fixed informative			
Day -2	0.180	0.162	-0.018	0.002	0.041	0.156
Day -1	0.110	0.110	0.000	0.001	0.035	0.135
Day 0	0.200	0.202	0.002	0.002	0.043	0.166
Day 1	0.180	0.187	0.007	0.002	0.041	0.160
Day 2	0.140	0.144	0.004	0.001	0.039	0.147
			Non-informative			
Day -2	0.180	0.176	-0.004	0.002	0.043	0.163
Day -1	0.110	0.110	0.000	0.001	0.037	0.139
Day 0	0.200	0.198	-0.002	0.002	0.044	0.170
Day 1	0.180	0.176	-0.004	0.002	0.042	0.163
Day 2	0.140	0.137	-0.003	0.001	0.040	0.150

Posterior distribution for the peak intensity

true	mean	bias	50% CR	90% CR
0.200	0.239	0.039	0.045	0.089

Posterior distribution for the peak intensity location

Day -2	Day -1	Day 0	Day 1	Day 2
0.000	0.007	0.536	0.427	0.030

Table B.10: Correctly specified prior parameters, 7 fertile window days, $n = 100$

day	true	mean	bias	MSE	Std.	CR
				Peak-intensity		
Day -4	0.059	0.062	0.003	0.001	0.030	0.114
Day -3	0.115	0.115	0.000	0.002	0.046	0.175
Day -2	0.196	0.184	-0.012	0.003	0.058	0.227
Day -1	0.288	0.257	-0.031	0.004	0.065	0.254
Day 0	0.340	0.289	-0.051	0.006	0.069	0.271
Day 1	0.264	0.232	-0.032	0.008	0.077	0.297
Day 2	0.104	0.124	0.020	0.006	0.069	0.257
				Fixed informative		
Day -4	0.059	0.027	-0.032	0.003	0.027	0.078
Day -3	0.115	0.106	-0.009	0.004	0.068	0.250
Day -2	0.196	0.191	-0.005	0.004	0.083	0.318
Day -1	0.288	0.283	-0.005	0.004	0.094	0.366
Day 0	0.340	0.333	-0.007	0.006	0.100	0.389
Day 1	0.264	0.263	-0.001	0.005	0.092	0.356
Day 2	0.104	0.087	-0.017	0.004	0.061	0.221
				Non-informative		
Day -4	0.059	0.093	0.034	0.005	0.071	0.256
Day -3	0.115	0.134	0.019	0.005	0.083	0.305
Day -2	0.196	0.180	-0.016	0.007	0.094	0.353
Day -1	0.288	0.250	-0.038	0.009	0.105	0.405
Day 0	0.340	0.290	-0.050	0.012	0.111	0.427
Day 1	0.264	0.237	-0.027	0.009	0.103	0.393
Day 2	0.104	0.128	0.024	0.006	0.081	0.297

Posterior distribution for the peak intensity

true	mean	bias	50% CR	90% CR
0.340	0.330	-0.010	0.072	0.139

Posterior distribution for the peak intensity location

Day -4	Day -3	Day -2	Day -1	Day 0	Day 1	Day 2
0.000	0.001	0.045	0.209	0.436	0.276	0.034

Table B.11: Correctly specified prior parameters, 7 fertile window days, $n = 200$

day	true	mean	bias	MSE	Std.	CR
				Peak-intensity		
Day -4	0.059	0.057	-0.002	0.000	0.021	0.081
Day -3	0.115	0.111	-0.004	0.001	0.034	0.131
Day -2	0.196	0.188	-0.008	0.002	0.046	0.179
Day -1	0.288	0.274	-0.014	0.002	0.054	0.211
Day 0	0.340	0.322	-0.018	0.002	0.057	0.225
Day 1	0.264	0.251	-0.013	0.003	0.059	0.231
Day 2	0.104	0.113	0.009	0.002	0.048	0.182
				Fixed informative		
Day -4	0.059	0.032	-0.027	0.002	0.027	0.082
Day -3	0.115	0.110	-0.005	0.003	0.055	0.205
Day -2	0.196	0.198	0.002	0.004	0.065	0.253
Day -1	0.288	0.290	0.002	0.004	0.073	0.286
Day 0	0.340	0.341	0.001	0.005	0.078	0.303
Day 1	0.264	0.265	0.001	0.004	0.072	0.277
Day 2	0.104	0.094	-0.010	0.003	0.052	0.190
				Non-informative		
Day -4	0.059	0.073	0.014	0.002	0.051	0.183
Day -3	0.115	0.123	0.008	0.004	0.061	0.229
Day -2	0.196	0.191	-0.005	0.005	0.071	0.272
Day -1	0.288	0.273	-0.015	0.006	0.078	0.305
Day 0	0.340	0.319	-0.021	0.007	0.082	0.321
Day 1	0.264	0.250	-0.014	0.005	0.077	0.296
Day 2	0.104	0.113	0.009	0.003	0.060	0.221

Posterior distribution for the peak intensity

true	mean	bias	50% CR	90% CR
0.340	0.335	-0.005	0.058	0.111

Posterior distribution for the peak
intensity location

Day -4	Day -3	Day -2	Day -1	Day 0	Day 1	Day 2
0.000	0.000	0.001	0.114	0.717	0.165	0.003

Table B.12: Correctly specified prior parameters, 7 fertile window days, $n = 300$

day	true	mean	bias	MSE	Std.	CR
			Peak-intensity			
Day -4	0.059	0.057	-0.002	0.000	0.017	0.067
Day -3	0.115	0.111	-0.004	0.000	0.028	0.109
Day -2	0.196	0.189	-0.007	0.001	0.039	0.151
Day -1	0.288	0.281	-0.007	0.001	0.047	0.184
Day 0	0.340	0.335	-0.005	0.001	0.051	0.200
Day 1	0.264	0.257	-0.007	0.002	0.048	0.190
Day 2	0.104	0.108	0.004	0.001	0.036	0.137
			Fixed informative			
Day -4	0.059	0.038	-0.021	0.002	0.028	0.088
Day -3	0.115	0.111	-0.004	0.002	0.048	0.179
Day -2	0.196	0.194	-0.002	0.003	0.055	0.215
Day -1	0.288	0.293	0.005	0.003	0.062	0.243
Day 0	0.340	0.350	0.010	0.004	0.066	0.257
Day 1	0.264	0.265	0.001	0.003	0.061	0.235
Day 2	0.104	0.094	-0.010	0.003	0.045	0.166
			Non-informative			
Day -4	0.059	0.067	0.008	0.002	0.042	0.151
Day -3	0.115	0.119	0.004	0.002	0.051	0.192
Day -2	0.196	0.188	-0.008	0.003	0.059	0.227
Day -1	0.288	0.282	-0.006	0.004	0.065	0.253
Day 0	0.340	0.336	-0.004	0.004	0.069	0.268
Day 1	0.264	0.255	-0.009	0.004	0.064	0.247
Day 2	0.104	0.107	0.003	0.003	0.049	0.184

Posterior distribution for the peak intensity

true	mean	bias	50% CR	90% CR
0.340	0.337	-0.003	0.051	0.100

Posterior distribution for the peak
intensity location

Day -4	Day -3	Day -2	Day -1	Day 0	Day 1	Day 2
0.000	0.000	0.000	0.059	0.844	0.096	0.001

Table B.13: Scaled and shifted prior parameters, 7 fertile window days, $n = 100$

day	true	mean	bias	MSE	Std.	CR
			Peak-intensity			
Day -2	0.235	0.238	0.003	0.003	0.067	0.265
Day -1	0.346	0.337	-0.009	0.004	0.079	0.310
Day 0	0.408	0.369	-0.039	0.005	0.077	0.303
Day 1	0.317	0.267	-0.050	0.007	0.070	0.273
Day 2	0.124	0.103	-0.021	0.002	0.044	0.166
Day 3	0.020	0.019	-0.001	0.000	0.016	0.051
Day 4	0.002	0.003	0.001	0.000	0.005	0.007
			Fixed informative			
Day -2	0.235	0.105	-0.130	0.025	0.061	0.214
Day -1	0.346	0.250	-0.096	0.016	0.099	0.382
Day 0	0.408	0.319	-0.089	0.014	0.104	0.405
Day 1	0.317	0.294	-0.023	0.007	0.098	0.381
Day 2	0.124	0.216	0.092	0.011	0.083	0.320
Day 3	0.020	0.130	0.110	0.014	0.068	0.256
Day 4	0.002	0.020	0.018	0.001	0.028	0.085
			Non-informative			
Day -2	0.235	0.217	-0.018	0.007	0.100	0.380
Day -1	0.346	0.309	-0.037	0.009	0.114	0.441
Day 0	0.408	0.339	-0.069	0.013	0.117	0.457
Day 1	0.317	0.266	-0.051	0.012	0.108	0.416
Day 2	0.124	0.137	0.013	0.004	0.083	0.308
Day 3	0.020	0.064	0.044	0.004	0.060	0.213
Day 4	0.002	0.043	0.041	0.003	0.050	0.173

Posterior distribution for the peak intensity

true	mean	bias	50% CR	90% CR
0.408	0.385	-0.023	0.088	0.169

Posterior distribution for the peak intensity location

Day -2	Day -1	Day 0	Day 1	Day 2	Day 3	Day 4
0.001	0.187	0.773	0.037	0.002	0.000	0.000

Table B.14: Scaled and shifted prior parameters, 7 fertile window days, $n = 200$

day	true	mean	bias	MSE	Std.	CR
			Peak-intensity			
Day -2	0.235	0.230	-0.005	0.001	0.049	0.192
Day -1	0.346	0.334	-0.012	0.002	0.060	0.234
Day 0	0.408	0.388	-0.020	0.002	0.063	0.248
Day 1	0.317	0.294	-0.023	0.003	0.055	0.216
Day 2	0.124	0.115	-0.009	0.001	0.032	0.124
Day 3	0.020	0.019	-0.001	0.000	0.010	0.030
Day 4	0.002	0.002	0.000	0.000	0.004	0.004
			Fixed informative			
Day -2	0.235	0.164	-0.071	0.011	0.063	0.238
Day -1	0.346	0.286	-0.060	0.009	0.078	0.301
Day 0	0.408	0.355	-0.053	0.007	0.081	0.318
Day 1	0.317	0.298	-0.019	0.005	0.075	0.292
Day 2	0.124	0.175	0.051	0.005	0.060	0.233
Day 3	0.020	0.087	0.067	0.006	0.047	0.175
Day 4	0.002	0.009	0.007	0.000	0.017	0.045
			Non-informative			
Day -2	0.235	0.224	-0.011	0.004	0.073	0.283
Day -1	0.346	0.321	-0.025	0.006	0.082	0.319
Day 0	0.408	0.375	-0.033	0.006	0.086	0.335
Day 1	0.317	0.289	-0.028	0.006	0.079	0.306
Day 2	0.124	0.129	0.005	0.003	0.060	0.228
Day 3	0.020	0.041	0.021	0.002	0.039	0.134
Day 4	0.002	0.023	0.021	0.001	0.031	0.103

Posterior distribution for the peak intensity

true	mean	bias	50% CR	90% CR
0.408	0.389	-0.019	0.070	0.135

Posterior distribution for the peak intensity location

Day -2	Day -1	Day 0	Day 1	Day 2	Day 3	Day 4
0.000	0.058	0.926	0.016	0.000	0.000	0.000

Table B.15: Scaled and shifted prior parameters, 7 fertile window days, $n = 300$

day	true	mean	bias	MSE	Std.	CR
				Peak-intensity		
Day -2	0.235	0.232	-0.003	0.001	0.042	0.166
Day -1	0.346	0.341	-0.005	0.002	0.052	0.205
Day 0	0.408	0.402	-0.006	0.002	0.056	0.221
Day 1	0.317	0.304	-0.013	0.002	0.049	0.192
Day 2	0.124	0.118	-0.006	0.000	0.027	0.107
Day 3	0.020	0.019	-0.001	0.000	0.008	0.023
Day 4	0.002	0.002	0.000	0.000	0.003	0.002
				Fixed informative		
Day -2	0.235	0.190	-0.045	0.006	0.058	0.222
Day -1	0.346	0.309	-0.037	0.005	0.066	0.257
Day 0	0.408	0.378	-0.030	0.005	0.070	0.274
Day 1	0.317	0.304	-0.013	0.003	0.064	0.251
Day 2	0.124	0.160	0.036	0.003	0.051	0.196
Day 3	0.020	0.067	0.047	0.003	0.037	0.136
Day 4	0.002	0.004	0.002	0.000	0.013	0.029
				Non-informative		
Day -2	0.235	0.229	-0.006	0.003	0.062	0.238
Day -1	0.346	0.335	-0.011	0.004	0.069	0.267
Day 0	0.408	0.395	-0.013	0.005	0.073	0.284
Day 1	0.317	0.300	-0.017	0.004	0.066	0.259
Day 2	0.124	0.126	0.002	0.002	0.051	0.195
Day 3	0.020	0.031	0.011	0.001	0.030	0.103
Day 4	0.002	0.016	0.014	0.000	0.024	0.076

Posterior distribution for the peak intensity

true	mean	bias	50% CR	90% CR
0.408	0.397	-0.011	0.064	0.124

Posterior distribution for the peak
intensity location

Day -2	Day -1	Day 0	Day 1	Day 2	Day 3	Day 4
0.000	0.032	0.962	0.006	0.000	0.000	0.000

Table B.16: Incorrect decay prior parameters, 7 fertile window days, $n = 100$

day	true	mean	bias	MSE	Std.	CR
			Peak-intensity			
Day -3	0.080	0.073	-0.007	0.002	0.033	0.124
Day -2	0.180	0.120	-0.060	0.007	0.045	0.172
Day -1	0.110	0.157	0.047	0.005	0.048	0.186
Day 0	0.200	0.187	-0.013	0.002	0.049	0.190
Day 1	0.180	0.167	-0.013	0.003	0.049	0.191
Day 2	0.140	0.117	-0.023	0.006	0.051	0.191
Day 3	0.070	0.062	-0.008	0.004	0.042	0.152
			Fixed informative			
Day -3	0.080	0.031	-0.049	0.004	0.028	0.086
Day -2	0.180	0.116	-0.064	0.007	0.064	0.238
Day -1	0.110	0.141	0.031	0.003	0.065	0.249
Day 0	0.200	0.210	0.010	0.003	0.074	0.286
Day 1	0.180	0.216	0.036	0.004	0.074	0.285
Day 2	0.140	0.172	0.032	0.004	0.068	0.261
Day 3	0.070	0.063	-0.007	0.003	0.047	0.166
			Non-informative			
Day -3	0.080	0.100	0.020	0.003	0.066	0.238
Day -2	0.180	0.148	-0.032	0.005	0.076	0.285
Day -1	0.110	0.126	0.016	0.004	0.072	0.268
Day 0	0.200	0.172	-0.028	0.006	0.081	0.305
Day 1	0.180	0.162	-0.018	0.005	0.079	0.299
Day 2	0.140	0.138	-0.002	0.004	0.074	0.275
Day 3	0.070	0.095	0.025	0.004	0.063	0.228

Posterior distribution for the peak intensity

true	mean	bias	50% CR	90% CR
0.200	0.239	0.039	0.056	0.108

Posterior distribution for the peak intensity location

Day -3	Day -2	Day -1	Day 0	Day 1	Day 2	Day 3
0.001	0.024	0.224	0.277	0.268	0.173	0.035

Table B.17: Incorrect decay prior parameters, 7 fertile window days, $n = 200$

day	true	mean	bias	MSE	Std.	CR
			Peak-intensity			
Day -3	0.080	0.067	-0.013	0.001	0.025	0.094
Day -2	0.180	0.126	-0.054	0.005	0.037	0.145
Day -1	0.110	0.158	0.048	0.004	0.039	0.150
Day 0	0.200	0.211	0.011	0.001	0.042	0.163
Day 1	0.180	0.193	0.013	0.001	0.042	0.162
Day 2	0.140	0.124	-0.016	0.003	0.042	0.160
Day 3	0.070	0.048	-0.022	0.002	0.029	0.105
			Fixed informative			
Day -3	0.080	0.044	-0.036	0.003	0.030	0.098
Day -2	0.180	0.152	-0.028	0.004	0.055	0.207
Day -1	0.110	0.130	0.020	0.002	0.050	0.190
Day 0	0.200	0.212	0.012	0.003	0.057	0.221
Day 1	0.180	0.202	0.022	0.003	0.056	0.215
Day 2	0.140	0.157	0.017	0.002	0.052	0.200
Day 3	0.070	0.061	-0.009	0.002	0.039	0.138
			Non-informative			
Day -3	0.080	0.084	0.004	0.002	0.048	0.173
Day -2	0.180	0.170	-0.010	0.003	0.059	0.225
Day -1	0.110	0.120	0.010	0.003	0.053	0.200
Day 0	0.200	0.194	-0.006	0.004	0.061	0.234
Day 1	0.180	0.173	-0.007	0.003	0.059	0.226
Day 2	0.140	0.135	-0.005	0.003	0.056	0.211
Day 3	0.070	0.079	0.009	0.002	0.046	0.168

Posterior distribution for the peak intensity

true	mean	bias	50% CR	90% CR
0.200	0.245	0.045	0.045	0.087

Posterior distribution for the peak intensity location

Day -3	Day -2	Day -1	Day 0	Day 1	Day 2	Day 3
0.000	0.000	0.077	0.440	0.415	0.067	0.002

Table B.18: Incorrect decay prior parameters, 7 fertile window days, $n = 300$

day	true	mean	bias	MSE	Std.	CR
				Peak-intensity		
Day -3	0.080	0.065	-0.015	0.001	0.022	0.084
Day -2	0.180	0.129	-0.051	0.004	0.034	0.132
Day -1	0.110	0.150	0.040	0.002	0.033	0.131
Day 0	0.200	0.216	0.016	0.001	0.038	0.147
Day 1	0.180	0.198	0.018	0.001	0.036	0.143
Day 2	0.140	0.129	-0.011	0.002	0.038	0.143
Day 3	0.070	0.046	-0.024	0.001	0.025	0.089
				Fixed informative		
Day -3	0.080	0.057	-0.023	0.002	0.032	0.109
Day -2	0.180	0.161	-0.019	0.002	0.047	0.178
Day -1	0.110	0.119	0.009	0.001	0.041	0.157
Day 0	0.200	0.213	0.013	0.002	0.048	0.187
Day 1	0.180	0.194	0.014	0.002	0.047	0.180
Day 2	0.140	0.150	0.010	0.001	0.044	0.167
Day 3	0.070	0.060	-0.010	0.001	0.034	0.121
				Non-informative		
Day -3	0.080	0.086	0.006	0.002	0.041	0.150
Day -2	0.180	0.173	-0.007	0.002	0.049	0.187
Day -1	0.110	0.111	0.001	0.002	0.043	0.165
Day 0	0.200	0.202	0.002	0.002	0.050	0.195
Day 1	0.180	0.173	-0.007	0.002	0.048	0.187
Day 2	0.140	0.135	-0.005	0.002	0.046	0.175
Day 3	0.070	0.073	0.003	0.001	0.039	0.139

Posterior distribution for the peak intensity

true	mean	bias	50% CR	90% CR
0.200	0.248	0.048	0.042	0.082

Posterior distribution for the peak
intensity location

Day -3	Day -2	Day -1	Day 0	Day 1	Day 2	Day 3
0.000	0.000	0.018	0.462	0.509	0.012	0.000

B.3 The day-specific probabilities model

The goal of this appendix is to fully characterize the Dunson and Stanford day-specific probabilities model. In its current state it tries to provide full detail of the derivations described in *Bayesian Inferences on Predictors of Conception Probabilities*.

B.3.1 Model specification

We wish to model the probability of a woman becoming pregnant for a given menstrual cycle as a function of her covariate status across the days of the cycle. Consider a study cohort and let us index

woman i,	$i = 1, \ldots, n,$
cycle j,	$j = 1, \ldots, n_i,$
day k,	$k = 1, \ldots, K,$

where day k refers to the k^{th} day out of a total of K days in the fertile window. Let us write day i, j, k as a shorthand for individual i, cycle j, and day k and similarly for cycle j, k. Then define

Y_{ij} an indicator of conception for woman i, cycle j,

V_{ijk} an indicator of conception for woman i, cycle j, day k,

X_{ijk} an indicator of intercourse for woman i, cycle j, day k.

Then writing $\boldsymbol{X}_{ij} = \big(X_{ij1}, \ldots, X_{ijK}\big)$, we observe that

$$\begin{aligned}
&\mathbb{P}\left(Y_{ij} = 1 \,\middle|\, \boldsymbol{X}_{ij},\ Y_{i1} = 0, \ldots, Y_{i,j-1} = 0\right) \\
&\quad = 1 - \mathbb{P}\left(Y_{ij} = 0 \,\middle|\, \boldsymbol{X}_{ij},\ Y_{i1} = 0, \ldots, Y_{i,j-1} = 0\right) \\
&\quad = 1 - \mathbb{P}\left(V_{ijk} = 0,\ k = 1, \ldots, K \,\middle|\, \boldsymbol{X}_{ij},\ Y_{i1} = 0, \ldots, Y_{i,j-1} = 0\right) \\
&\quad = 1 - \prod_{k=1}^{K} \mathbb{P}\left(V_{ijk} = 0 \,\middle|\, X_{ijk},\ Y_{i1} = 0, \ldots, Y_{i,j-1} = 0,\ V_{ij1} = 0, \ldots, V_{i,k-1} = 0\right) \\
&\quad = 1 - \prod_{k=1}^{K} \left\{1 - \mathbb{P}\left(V_{ijk} = 1 \,\middle|\, X_{ijk},\ Y_{i1} = 0, \ldots, Y_{i,j-1} = 0,\ V_{ij1} = 0, \ldots, V_{i,k-1} = 0\right)\right\} \\
&\quad = 1 - \prod_{k=1}^{K} \left\{1 - X_{ijk}\, \mathbb{P}\left(V_{ijk} = 1 \,\middle|\, Y_{i1} = 0, \ldots, Y_{i,j-1} = 0,\ V_{ij1} = 0, \ldots, V_{i,k-1} = 0\right)\right\} \\
&\quad = 1 - \prod_{k=1}^{K} \left\{1 - \mathbb{P}\left(V_{ijk} = 1 \,\middle|\, Y_{i1} = 0, \ldots, Y_{i,j-1} = 0,\ V_{ij1} = 0, \ldots, V_{i,k-1} = 0\right)\right\}^{X_{ijk}}.
\end{aligned}$$

With this result in mind, we now consider the Dunson and Stanford day-specific probabilities model. Using the same indexing scheme as above, define

$\boldsymbol{u}_{ijk}$ a covariate vector of length q for woman i, cycle j, day k,

$\boldsymbol{\beta}$ a vector of length q of regression coefficients,

ξ_i woman-specific random effect.

Then writing $\boldsymbol{U}_{ij} = \big(\boldsymbol{u}'_{ijk}, \ldots, \boldsymbol{u}'_{ijk}\big)'$, Dunson and Stanford propose the model:

$$\begin{aligned}
\mathbb{P}\left(Y_{ij} = 1 \,\middle|\, \xi_i, \boldsymbol{X}_{ij}, \boldsymbol{U}_{ij}\right) &= 1 - \prod_{k=1}^{K} (1 - \lambda_{ijk})^{X_{ijk}}, \\
\lambda_{ijk} &= 1 - \exp\left\{-\xi_i \exp\left(\boldsymbol{u}'_{ijk}\boldsymbol{\beta}\right)\right\}, \\
\xi_i &\sim \text{Gamma}\,(\phi, \phi). \qquad \text{(B.12)}
\end{aligned}$$

From our previous derivation, we see that we may interpret λ_{ijk} as the day-specific probability of conception in cycle j from couple i given that conception has not already occured, or in the language of Dunson and Stanford, given intercourse only on day k.

Delving further, we see that λ_{ijk} is strictly increasing in $u_{ijkh}\,\beta_h$, where we are denoting u_{ijkh} to be the h^{th} term in $\boldsymbol{u}_{ijk}$ and similarly for β_h. When $\beta_h = 0$ then the h^{th} covariate has no effect on the day-specific probability of conception.

λ_{ijk} is also strictly increasing in ξ_i which as Dunson and Stanford suggest may be interpreted as a woman-specific random effect. The authors state that specifying the distribution of the ξ_i with a common parameters prevents nonidentifiability between $\mathbb{E}\left[\xi_i\right]$ and the day-specific parameters. Since Var $\left[\xi_i\right] = 1/\phi$ it follows that ϕ may be interpreted as a measure of variability across women.

B.3.1.1 Computation consideration

As an aside, we note that it may be more computationally convenient to calculate

$$\begin{aligned}
&\mathbb{P}\left(Y_{ij} = 1 \,\middle|\, \xi_i, \boldsymbol{X}_{ij}, \boldsymbol{U}_{ij}\right) \\
&\quad = 1 - \prod_{k=1}^{K} \left(1 - \lambda_{ijk}\right)^{X_{ijk}} \\
&\quad = 1 - \prod_{k=1}^{K} \left[\exp\left\{-\xi_i \exp\left(\boldsymbol{u}'_{ijk}\boldsymbol{\beta}\right)\right\}\right]^{X_{ijk}} \\
&\quad = 1 - \prod_{k=1}^{K} \exp\left\{-X_{ijk}\xi_i \exp\left(\boldsymbol{u}'_{ijk}\boldsymbol{\beta}\right)\right\} \\
&\quad = 1 - \exp\left\{-\sum_{k=1}^{K} X_{ijk}\xi_i \exp\left(\boldsymbol{u}'_{ijk}\boldsymbol{\beta}\right)\right\}.
\end{aligned}$$

B.3.2 Marginal probability of conception

The marginal probability of conception, obtained by integrating out the couple-specific frailty ξ_i, has form as follows.

$$\begin{aligned}
&\mathbb{P}\left(Y_{ij} = 1 \,|\, \boldsymbol{X}_{ij}, \boldsymbol{U}_{ij}\right) \\
&\quad = \int_0^\infty \mathbb{P}\left(Y_{ij}, \xi_i \,\middle|\, \boldsymbol{X}_{ij}, \boldsymbol{U}_{ij}\right) d\xi_i
\end{aligned}$$

$$
\begin{aligned}
&= \int_0^\infty \mathbb{P}\left(Y_{ij}, \xi_i \,|\, \boldsymbol{X}_{ij}, \boldsymbol{U}_{ij}\right) \mathcal{G}(\xi_i;\, \phi, \phi)\, d\xi_i \\
&= \int_0^\infty \left[1 - \prod_{k=1}^{K} (1-\lambda_{ijk})^{X_{ijk}}\right] \mathcal{G}(\xi_i;\, \phi, \phi)\, d\xi_i \\
&= 1 - \int_0^\infty \prod_{k=1}^{K} (1-\lambda_{ijk})^{X_{ijk}}\, \mathcal{G}(\xi_i;\, \phi, \phi)\, d\xi_i \\
&= 1 - \int_0^\infty \prod_{k=1}^{K} \left[\exp\left\{-\xi_i \exp\left(\boldsymbol{u}'_{ijk}\boldsymbol{\beta}\right)\right\}\right]^{X_{ijk}} \mathcal{G}(\xi_i;\, \phi, \phi)\, d\xi_i \\
&= 1 - \int_0^\infty \prod_{k=1}^{K} \exp\left\{ -\xi_i X_{ijk} \exp\left(\boldsymbol{u}'_{ijk}\boldsymbol{\beta}\right)\right\} \mathcal{G}(\xi_i;\, \phi, \phi)\, d\xi_i \\
&= 1 - \int_0^\infty \exp\left\{-\xi_i \sum_{k=1}^{K} X_{ijk} \exp\left(\boldsymbol{u}'_{ijk}\boldsymbol{\beta}\right)\right\} \mathcal{G}(\xi_i;\, \phi, \phi)\, d\xi_i \\
&= 1 - \left[\frac{\phi}{\phi + \sum_{k=1}^{K} X_{ijk} \exp\left(\boldsymbol{u}'_{ijk}\boldsymbol{\beta}\right)}\right]^{\phi},
\end{aligned}
$$

since

$$
\begin{aligned}
&\int_0^\infty \exp\left\{-\xi_i \sum_{k=1}^{K} X_{ijk} \exp\left(\boldsymbol{u}'_{ijk}\boldsymbol{\beta}\right)\right\} \mathcal{G}(\xi_i;\, \phi, \phi)\, d\xi_i \\
&\quad= \int_0^\infty \exp\left\{-\xi_i \sum_{k=1}^{K} X_{ijk} \exp\left(\boldsymbol{u}'_{ijk}\boldsymbol{\beta}\right)\right\} \frac{\phi^\phi}{\Gamma(\phi)}\, \xi_i^{\phi-1} d\xi_i \\
&\quad= \int_0^\infty \frac{\phi^\phi}{\Gamma(\phi)}\, \xi_i^{\phi-1} \exp\left\{-\xi_i \left[\phi + \sum_{k=1}^{K} X_{ijk} \exp\left(\boldsymbol{u}'_{ijk}\boldsymbol{\beta}\right)\right]\right\} d\xi_i \\
&\quad= \left[\frac{\phi}{\phi + \sum_{k=1}^{K} X_{ijk} \exp\left(\boldsymbol{u}'_{ijk}\boldsymbol{\beta}\right)}\right]^{\phi} \int_0^\infty \frac{\left[\phi + \sum_{k=1}^{K} X_{ijk} \exp\left(\boldsymbol{u}'_{ijk}\boldsymbol{\beta}\right)\right]^{\phi}}{\Gamma(\phi)} \\
&\qquad\qquad \times \xi_i^{\phi-1} \exp\left\{-\xi_i \left[\phi + \sum_{k=1}^{K} X_{ijk} \exp\left(\boldsymbol{u}'_{ijk}\boldsymbol{\beta}\right)\right]^{\phi}\right\} d\xi_i,
\end{aligned}
$$

and the function inside the integral is a gamma density function.

B.3.2.1 Day-specific marginal probability of conception

Dunson and Stanford also point out the following remarkable result. The marginal day-specific probability of conception in a cycle with intercourse only on day k and with predictors $\boldsymbol{u}$ is given by

$$\mathbb{P}\left(Y=1 \,|\, \boldsymbol{u}\right) = 1 - \left(\frac{\phi}{\phi + \exp(\boldsymbol{u}'\boldsymbol{\beta})}\right)^{\phi},$$

which is in the form of the Aranda-Ordaz generalized linear model, and reduces to a logistic regression model for $\phi = 1$.

B.3.3 Prior specification

Define

$\mathcal{G}_{\mathcal{A}_h}(\cdot)$ density function of a gamma distribution truncated to the region $\mathcal{A}_h \subset (0, \infty)$,

γ_h $\exp(\beta_h)$.

Then the Dunson and Stanford model chooses priors of the form

$$\pi\left(\boldsymbol{\gamma}\right) = \prod_{h=1}^{q} \Big\{ p_h \mathbb{1}\left(\gamma_h = 1\right) \;+\; (1-p_h)\,\mathbb{1}\left(\gamma_h \neq 1\right) \mathcal{G}_{\mathcal{A}_h}\big(\gamma_h;\; a_h, b_h\big) \Big\},$$

$$\pi(\phi) = \mathcal{G}\left(\phi;\; c_1, c_2\right).$$

where

p_h prior probability that $\gamma_h = 1$, a hyperparameter,

a_h, b_h shape and rate hyperparameters for gamma distribution of γ_h,

c_1, c_2 shape and rate hyperparameters for gamma distribution of ϕ.

Values of $\gamma_h = 1$ correspond to $\beta_h = 0$ and the h^{th} predictor in $\boldsymbol{u}_{ijk}$ being dropped from the model. Thus assigning the prior for each of the γ_h to be a mixture distribution between a point mass at one and a gamma distribution allows the model to drop terms from the regression component with nonzero probability.

Typical constraints for the γ_h are $\mathbb{R}^+$, $(0,1)$, and $(1,\infty)$ which correspond to no constraint, a negative effect on probability of conception, and a positive effect on probability of conception, respectively. Thus a priori knowledge of the direction of association of the predictor variables can be incorporated into the model to decrease posterior uncertainty.

B.3.3.1 Monotone effects

Consider a model where the list of covariates includes an ordered categorical variable with types $1, \ldots, t$. Let $\boldsymbol{s}_{ijk} = \big(s_{ijk,2}, \ldots, s_{ijk,t}\big)$ be a vector of length $(t-1)$ for each day i, j, k where

$$\begin{aligned}
s_{ijk,2} &= I\Big(\text{categorical variable for day } i,j,k \text{ is type 2}\Big) \\
s_{ijk,3} &= I\Big(\text{categorical variable for day } i,j,k \text{ is type 2 or 3}\Big) \\
\vdots \quad &\ \vdots \qquad\qquad\qquad\qquad \vdots \\
s_{ijk,t} &= I\Big(\text{categorical variable for day } i,j,k \text{ is type 2 or 3 or} \ldots \text{or } t\Big).
\end{aligned}$$

Next, let us partition each covariate vector $\boldsymbol{u}_{ijk} = \big(\boldsymbol{r}_{ijk},\, \boldsymbol{s}_{ijk}\big)$ so that $\boldsymbol{r}_{ijk}$ is a vector of the remaining covariate terms. Furthermore let $\boldsymbol{\beta} = \big(\boldsymbol{\tau},\, \boldsymbol{\alpha}\big)$ be the corresponding partition of covariate coefficients where $\boldsymbol{\alpha} = \big(\alpha_2, \ldots, \alpha_t\big)$. Then for person i, cycle j, and day k with categorical variable type d where $d \in \{1, \ldots, t\}$, then

$$\begin{aligned}
\lambda_{ijk} &= 1 - \exp\big\{-\xi_i \exp\big(\boldsymbol{u}'_{ijk}\boldsymbol{\beta}\big)\big\} \\
&= 1 - \exp\big\{-\xi_i \exp\big(\boldsymbol{r}'_{ijk}\boldsymbol{\tau} + \boldsymbol{s}'_{ijk}\boldsymbol{\alpha}\big)\big\} \\
&= 1 - \mathbb{1}\ \big(d = 1\big)\ \exp\big\{-\xi_i \exp\big(\boldsymbol{r}'_{ijk}\boldsymbol{\tau}\big)\big\} \\
&\qquad - \mathbb{1}\ \big(d \geq 2\big)\ \exp\left\{-\xi_i \exp\left(\boldsymbol{r}'_{ijk}\boldsymbol{\tau} + \sum_{m=2}^{d} \alpha_m\right)\right\}
\end{aligned}$$

From this form we can see that when $\alpha_m \geq 0,\ m = 2, \ldots, t$ then λ_{ijk} is nondecreasing in m. It follows that a monotone increasing categorical variable can be created by coding the variable in the format as described above, and constraining the corresponding parameters of γ_h to be greater than or equal to one (corresonding to $\beta_h \geq 0$ for each of the corresonding h). Similarly, a monotone decreasing categorical variable can be created by coding the variable as described above, and constraining the corresponding parameters of γ_h to be less than or equal to one.

B.3.4 Posterior computation

Express the data augmentation model as

$$Y_{ij} = \mathbb{1}\left(\sum_{k=1}^{K} X_{ijk} Z_{ijk} > 0\right),$$

$$Z_{ijk} \sim \text{Poisson}\left(\xi_i \exp\left(\boldsymbol{u}'_{ijk}\boldsymbol{\beta}\right)\right), \qquad k = 1, \dots, K. \tag{B.13}$$

Let us further define $W_{ijk} = X_{ijk} Z_{ijk}$ for all i, j, k.

B.3.4.1 Verifying the equivalence of the data augmentation model

Under (B.13), $Y_{ij} = 0$ if and only if $W_{ij1}, \dots, W_{ijK}$ are identically 0. It follows that

$$\begin{aligned}
&\mathbb{P}\left(Y_{ij} = 0 \mid \xi_i, \boldsymbol{X}_{ij}, \boldsymbol{U}_{ij}\right) \\
&\quad = \prod_{k:\, X_{ijk}=1} \mathbb{P}\left(W_{ijk} = 0 \mid \xi_i, \boldsymbol{u}_{ijk}\right) \\
&\quad = \prod_{k=1}^{K} \left[\mathbb{P}\left(W_{ijk} = 0 \mid \xi_i, \boldsymbol{u}_{ijk}\right)\right]^{X_{ijk}} \\
&\quad = \prod_{k=1}^{K} \left[\exp\left\{\xi_i \exp\left(\boldsymbol{u}'_{ijk}\boldsymbol{\beta}\right)\right\}\right]^{X_{ijk}} \\
&\quad = \prod_{k=1}^{K} (1 - \lambda_{ijk})^{X_{ijk}}.
\end{aligned}$$

which is the model in (B.12).

B.3.4.2 The full likelihood

Let $\boldsymbol{Y}$ be a random variable representing all of the potential pregnancy indicators Y_{ij}, let $\boldsymbol{W}$ be a random variable representing all of the latent variables W_{ijk}, and let ξ be a random variable representing all of the woman-specific random effects ξ_i. Then

$$
\begin{aligned}
&\pi\left(\boldsymbol{Y}, \boldsymbol{W}, \gamma, \boldsymbol{\xi}, \phi \,|\, \text{data}\right) \\
&\quad = \pi\left(\boldsymbol{Y} \,|\, \boldsymbol{W}, \boldsymbol{\gamma}, \boldsymbol{\xi}, \phi, \text{data}\right) \pi\left(\boldsymbol{W} \,|\, \boldsymbol{\gamma}, \boldsymbol{\xi}, \phi, \text{data}\right) \pi\left(\boldsymbol{\xi} \,|\, \boldsymbol{\gamma}, \phi, \text{data}\right) \pi\left(\boldsymbol{\gamma} \,|\, \phi, \text{data}\right) \pi\left(\phi \,|\, \text{data}\right) \\
&\quad = \pi\left(\boldsymbol{Y} \,|\, \boldsymbol{W}\right) \pi\left(\boldsymbol{W} \,|\, \boldsymbol{\gamma}, \boldsymbol{\xi}, \text{data}\right) \pi\left(\boldsymbol{\xi} \,|\, \phi\right) \pi\left(\boldsymbol{\gamma}\right) \pi\left(\phi\right) \\
&\quad = \left(\prod_{i,j} \pi\left(Y_{ij} \,|\, \boldsymbol{W}_{ij}\right)\right)\left(\prod_{i,j,k\colon X_{ijk}=1} \pi\left(W_{ijk} \,|\, \boldsymbol{\gamma}, \boldsymbol{\xi}\right)\right)\left(\prod_{i=1}^{n} \pi\left(\xi_i \,|\, \phi\right)\right)\left(\prod_{\ell=1}^{q} \pi\left(\gamma_h\right)\right) \pi\left(\phi\right) \\
&\quad = \left\{\prod_{i,j}\left[\mathbb{1}\left(\sum_{k=1}^{K} W_{ijk} > 0\right) Y_{ij} + \mathbb{1}\left(\sum_{k=1}^{K} W_{ijk} = 0\right)\left(1 - Y_{ij}\right)\right]\right\} \\
&\qquad \times \left(\prod_{i,j,k\colon X_{ijk}=1} \frac{1}{W_{ijk}!}\left[\xi_i \exp\left(\sum_{\ell=1}^{q} u_{ijk\ell} \log \gamma_\ell\right)\right]^{W_{ijk}} \exp\left\{-\xi_i \exp\left(\sum_{\ell=1}^{q} u_{ijk\ell} \log \gamma_\ell\right)\right\}\right) \\
&\qquad \times \left(\prod_{i=1}^{n} \frac{\phi^{\phi}}{\Gamma(\phi)}\, \xi_i^{\phi-1} \exp\left(-\phi \xi_i\right)\right) \\
&\qquad \times \left(\prod_{\ell=1}^{q}\left[p_h \mathbb{1}\left(\gamma_h = 1\right) \;+\; \left(1 - p_h\right) \mathbb{1}\left(\gamma_h \neq 1\right) \mathcal{G}_{\mathcal{A}_h}(\gamma_h;\, a_h, b_h)\right]\right) \\
&\qquad \times \frac{c_2^{c_1}}{\Gamma(c_1)}\, \phi^{c_1-1} \exp\left(-c_2 \phi\right).
\end{aligned}
$$

B.3.4.3 The full conditional distributions

The full conditional distribution for W

Writing $\boldsymbol{W}_{ij} = \left(W_{ij1}, \ldots, W_{ijK}\right)$ and letting $\boldsymbol{m} = \left(m_1, \ldots, m_K\right)$ be a vector of realized outcomes for $\boldsymbol{W}_{ij}$, we see first that for $Y_{ij} = 0$ we have

$$
\mathbb{P}\left(\boldsymbol{W}_{ij} = \boldsymbol{m} \;\middle|\; Y_{ij} = 0,\, \boldsymbol{\beta}, \phi, \boldsymbol{\xi}, \text{data}\right) \;=\; \begin{cases} 1, & \boldsymbol{m} = \boldsymbol{0}, \\[2ex] 0, & \text{else.} \end{cases}
$$

Next, for $Y_{ij} = 1$ we have

$$
\begin{aligned}
&\mathbb{P}\left(\boldsymbol{W}_{ij} = \boldsymbol{m} \;\middle|\; Y_{ij} = 1,\, \boldsymbol{\beta}, \phi, \boldsymbol{\xi}, \text{data}\right) \\
&\quad = \sum_{s=0}^{\infty} \mathbb{P}\left(\boldsymbol{W}_{ij} = \boldsymbol{m},\; \textstyle\sum_k W_{ijk} = s \;\middle|\; Y_{ij} = 1,\, \boldsymbol{\beta}, \phi, \boldsymbol{\xi}, \text{data}\right)
\end{aligned}
$$

$$
\begin{aligned}
&= \mathbb{P}\left(\boldsymbol{W}_{ij} = \boldsymbol{m},\ \textstyle\sum_k W_{ijk} = \sum_k m_k \,\middle|\, Y_{ij} = 1,\ \boldsymbol{\beta}, \phi, \boldsymbol{\xi}, \text{data}\right) \\
&= \mathbb{P}\left(\boldsymbol{W}_{ij} = \boldsymbol{m} \,\middle|\, \textstyle\sum_k W_{ijk} = \sum_k m_k,\ Y_{ij} = 1,\ \boldsymbol{\beta}, \phi, \boldsymbol{\xi}, \text{data}\right) \\
&\qquad \times \mathbb{P}\left(\textstyle\sum_k W_{ijk} = \sum_k m_k \,\middle|\, Y_{ij} = 1,\ \boldsymbol{\beta}, \phi, \boldsymbol{\xi}, \text{data}\right).
\end{aligned}
$$

Furthermore,

$$
\begin{aligned}
&\pi\left(\sum_{k=1}^{K} W_{ijk} \,\middle|\, Y_{ij} = 1,\ \boldsymbol{\beta}, \phi, \boldsymbol{\xi}, \text{data}\right) \\
&\qquad = \ \pi\left(\sum_{k=1}^{K} W_{ijk} \,\middle|\, \sum_{k=1}^{K} W_{ijk} \geq 1,\ \boldsymbol{\beta}, \phi, \boldsymbol{\xi}, \text{data}\right) \\
&\qquad \sim \ \text{Poisson}\left(\xi_i \sum_{k:\ X_{ijk}=1} \exp\left(\boldsymbol{u}'_{ijk}\boldsymbol{\beta}\right)\right) \ \text{truncated so that} \ \sum_{k=1}^{K} W_{ijk} \geq 1,
\end{aligned}
$$

and

$$
\begin{aligned}
&\pi\left(\boldsymbol{W}_{ij} \,\middle|\, \sum_{k=1}^{K} W_{ijk},\ Y_{ij} = 1,\ \boldsymbol{\beta}, \phi, \boldsymbol{\xi}, \text{data}\right) \\
&\qquad \sim \ \text{Multinomial}\left(\sum_{k=1}^{K} W_{ijk};\ \frac{X_{ij1}\xi_i \exp\left(\boldsymbol{u}'_{ij1}\boldsymbol{\beta}\right)}{\xi_i \displaystyle\sum_{k:\ X_{ijk}=1} \exp\left(\boldsymbol{u}'_{ijk}\boldsymbol{\beta}\right)},\ \ldots,\ \frac{X_{ijK}\xi_i \exp\left(\boldsymbol{u}'_{ijK}\boldsymbol{\beta}\right)}{\xi_i \displaystyle\sum_{k:\ X_{ijk}=1} \exp\left(\boldsymbol{u}'_{ijk}\boldsymbol{\beta}\right)}\right).
\end{aligned}
$$

The full conditional distribution for γ_h

Define the following terms which will be of use in the following derivation. Denote

$$
\begin{array}{ll}
\widetilde{a}_h & a_h + \sum_{i,j,k} u_{ijkh} W_{ijk}, \\[2ex]
\widetilde{b}_h & b_h + \displaystyle\sum_{\substack{i,j,k\colon X_{ijk}=1, \\ u_{ijkh}=1}} \xi_i \prod_{\ell \neq h} \gamma_\ell^{u_{ijk\ell}}, \\[4ex]
d_1 & p_h \exp\left\{ -(\widetilde{b}_h - b_h) \right\}, \\[2ex]
d_2 & (1-p_h)\, \dfrac{C(a_h, b_h) \int_{\mathcal{A}_h} \mathcal{G}(\gamma;\, \widetilde{a}_h, \widetilde{b}_h)\, d\gamma}{C(\widetilde{a}_h, \widetilde{b}_h) \int_{\mathcal{A}_h} \mathcal{G}(\gamma;\, a_h, b_h)\, d\gamma}, \\[2ex]
\widetilde{p}_h & \dfrac{d_1}{d_1 + d_2}.
\end{array}
$$

Then for the case when the explanetory variables are all categorical, we have

$$
\begin{aligned}
&\pi\left(\gamma_h \,|\, \boldsymbol{\gamma}_{(-h)}, \phi, \boldsymbol{\xi}, \boldsymbol{W}, \text{data}\right) \\
&\qquad \propto \pi\left(\boldsymbol{W} \mid \boldsymbol{\xi}, \gamma, \text{data}\right) \pi\left(\gamma_h\right) \\
&\qquad = \left(\prod_{i=1}^{n} \prod_{j=1}^{n_i} \prod_{k\colon X_{ijk}=1} \pi\left(W_{ijk} \,|\, \boldsymbol{\xi}_i, \gamma, \text{data}\right) \right) \pi\left(\gamma_h\right) \\
&\qquad \propto \left(\prod_{i=1}^{n} \prod_{j=1}^{n_i} \prod_{k\colon X_{ijk}=1} \Big[\exp\left(u_{ijkh} \log \gamma_h\right) \Big]^{W_{ijk}} \right. \\
&\qquad\qquad\quad \left. \times \exp\left\{ -\xi_i \exp\left(\sum_{\ell=1}^{q} u_{ijk\ell} \log \gamma_\ell \right) \right\} \right) \pi\left(\gamma_h\right) \\
&\qquad = \left(\prod_{i=1}^{n} \prod_{j=1}^{n_i} \prod_{k\colon X_{ijk}=1} \gamma_h^{u_{ijkh} W_{ijk}} \exp\left\{ -\xi_i \prod_{\ell=1}^{q} \gamma_\ell^{u_{ijk\ell}} \right\} \right) \pi\left(\gamma_h\right) \\
&\qquad = \gamma_h^{\sum_{i,j,k} u_{ijkh} W_{ijk}} \exp\left\{ - \sum_{i,j,k\colon X_{ijk}=1} \xi_i \prod_{\ell=1}^{q} \gamma_\ell^{u_{ijk\ell}} \right\} \pi\left(\gamma_h\right) \\
&\qquad = \gamma_h^{\sum_{i,j,k} u_{ijkh} W_{ijk}} \exp\left\{ - \sum_{\substack{i,j,k\colon X_{ijk}=1, \\ u_{ijkh}=0}} \xi_i \prod_{\ell \neq h} \gamma_\ell^{u_{ijk\ell}} - \gamma_h \sum_{\substack{i,j,k\colon X_{ijk}=1, \\ u_{ijkh}=1}} \xi_i \prod_{\ell \neq h} \gamma_\ell^{u_{ijk\ell}} \right\} \pi\left(\gamma_h\right)
\end{aligned}
$$

$$\propto \gamma_h^{\sum_{i,j,k} u_{ijkh} W_{ijk}} \exp\left\{ -\gamma_h \sum_{\substack{i,j,k:\, X_{ijk}=1,\\ u_{ijkh}=1}} \xi_i \prod_{\ell \neq h} \gamma_\ell^{u_{ijk\ell}} \right\} \pi\left(\gamma_h\right)$$

$$= \gamma_h^{\sum_{i,j,k} u_{ijkh} W_{ijk}} \exp\left\{ -\gamma_h \sum_{\substack{i,j,k:\, X_{ijk}=1,\\ u_{ijkh}=1}} \xi_i \prod_{\ell \neq h} \gamma_\ell^{u_{ijk\ell}} \right\}$$

$$\times \left[p_h \mathbb{1}\left(\gamma_h = 1\right) \; + \; (1 - p_h)\, \mathbb{1}\left(\gamma_h \neq 1\right) \mathcal{G}_{\mathcal{A}_h}(\gamma_h;\, a_h, b_h) \right]$$

$$= p_h \, \mathbb{1}\left(\gamma_h = 1\right) \, \exp\left\{ - \sum_{\substack{i,j,k:\, X_{ijk}=1,\\ u_{ijkh}=1}} \xi_i \prod_{\ell \neq h} \gamma_\ell^{u_{ijk\ell}} \right\}$$

$$+ \; (1 - p_h)\, \mathbb{1}\left(\gamma_h \neq 1\right) \gamma_h^{\sum_{i,j,k} u_{ijkh} W_{ijk}}$$

$$\times \exp\left\{ -\gamma_h \sum_{\substack{i,j,k:\, X_{ijk}=1,\\ u_{ijkh}=1}} \xi_i \prod_{\ell \neq h} \gamma_\ell^{u_{ijk\ell}} \right\} \mathcal{G}_{\mathcal{A}_h}(\gamma_h;\, a_h, b_h)$$

$$= p_h \, \mathbb{1}\left(\gamma_h = 1\right) \, \exp\left\{ - \sum_{\substack{i,j,k:\, X_{ijk}=1,\\ u_{ijkh}=1}} \xi_i \prod_{\ell \neq h} \gamma_\ell^{u_{ijk\ell}} \right\}$$

$$+ \; (1 - p_h)\, \mathbb{1}\left(\gamma_h \neq 1\right) \frac{C(a_h, b_h)}{\int_{\mathcal{A}_h} \mathcal{G}(\gamma;\, a_h, b_h)\, d\gamma} \, \gamma_h^{a_h + \sum_{i,j,k} u_{ijkh} W_{ijk} - 1}$$

$$\times \; \exp\left\{ -\gamma_h \left[b_h + \sum_{\substack{i,j,k:\, X_{ijk}=1,\\ u_{ijkh}=1}} \xi_i \prod_{\ell \neq h} \gamma_\ell^{u_{ijk\ell}} \right] \right\}$$

$$= p_h \, \mathbb{1}\left(\gamma_h = 1\right) \, \exp\left\{ -(\widetilde{b}_h - b_h) \right\}$$

$$+ (1 - p_h)\, \mathbb{1}\left(\gamma_h \neq 1\right) \frac{C(a_h, b_h)}{\int_{\mathcal{A}_h} \mathcal{G}(\gamma;\, a_h, b_h)\, d\gamma} \, \gamma_h^{\widetilde{a}_h - 1} \, \exp\left\{ -\widetilde{b}_h \gamma_h \right\}$$

$$= p_h \, \mathbb{1}\left(\gamma_h = 1\right) \, \exp\left\{ -(\widetilde{b}_h - b_h) \right\} \; + \; (1 - p_h)\, \mathbb{1}\left(\gamma_h \neq 1\right)$$

$$\times \; \frac{C(a_h, b_h) \int_{\mathcal{A}_h} \mathcal{G}(\gamma;\, \widetilde{a}_h, \widetilde{b}_h)\, d\gamma}{C(\widetilde{a}_h, \widetilde{b}_h) \int_{\mathcal{A}_h} \mathcal{G}(\gamma;\, a_h, b_h)\, d\gamma} \, \frac{C(\widetilde{a}_h, \widetilde{b}_h)}{\int_{\mathcal{A}_h} \mathcal{G}(\gamma;\, \widetilde{a}_h, \widetilde{b}_h)\, d\gamma} \gamma_h^{\widetilde{a}_h - 1} \, \exp\left\{ -\widetilde{b}_h \gamma_h \right\}$$

$$= p_h \, \mathbb{1}\left(\gamma_h = 1\right) \, \exp\left\{ -(\widetilde{b}_h - b_h) \right\}$$

$$
\begin{aligned}
&\qquad + (1-p_h)\,\mathbb{1}\left(\gamma_h \neq 1\right) \frac{C(a_h,b_h)\int_{\mathcal{A}_h}\mathcal{G}(\gamma;\,\widetilde{a}_h,\widetilde{b}_h)\,d\gamma}{C(\widetilde{a}_h,\widetilde{b}_h)\int_{\mathcal{A}_h}\mathcal{G}(\gamma;\,a_h,b_h)\,d\gamma}\,\mathcal{G}_{\mathcal{A}_h}\big(\gamma;\,\widetilde{a}_h,\widetilde{b}_h\big) \\
&= d_1\mathbb{1}\left(\gamma_h = 1\right) \;+\; d_2\mathbb{1}\left(\gamma_h \neq 1\right)\mathcal{G}_{\mathcal{A}_h}\big(\gamma;\,\widetilde{a}_h,\widetilde{b}_h\big) \\
&\propto \frac{d_1}{d_1+d_2}\mathbb{1}\left(\gamma_h = 1\right) \;+\; \frac{d_2}{d_1+d_2}\mathbb{1}\left(\gamma_h \neq 1\right)\mathcal{G}_{\mathcal{A}_h}\big(\gamma;\,\widetilde{a}_h,\widetilde{b}_h\big) \\
&= \widetilde{p}_h\mathbb{1}\left(\gamma_h = 1\right) \;+\; (1-\widetilde{p}_h)\,\mathbb{1}\left(\gamma_h \neq 1\right)\mathcal{G}_{\mathcal{A}_h}\big(\gamma;\,\widetilde{a}_h,\widetilde{b}_h\big).
\end{aligned}
$$

The full conditional distribution for ξ_i

$$
\begin{aligned}
&\pi\left(\boldsymbol{\xi}_i \,\middle|\, \boldsymbol{\beta},\phi,\boldsymbol{W},\text{data}\right) \\
&\qquad\propto \pi\left(\boldsymbol{W}_i \,\middle|\, \boldsymbol{\beta},\xi_i,\text{data}\right)\,\pi\left(\xi_i \,|\, \phi,\text{data}\right) \\
&\qquad= \left(\prod_{j,k:\,X_{ijk}=1}\pi\left(\boldsymbol{W}_{ijk} \,\middle|\, \boldsymbol{\beta},\xi_i,\text{data}\right)\right)\pi\left(\xi_i \,|\, \phi,\text{data}\right) \\
&\qquad\propto \left(\prod_{j,k:\,X_{ijk}=1}\xi_i^{W_{ijk}}\exp\left\{-\,\xi_i\exp\left(\boldsymbol{u}_{ijk}'\boldsymbol{\beta}\right)\right\}\right)\xi_i^{\phi-1}\exp\left\{-\,\phi\xi_i\right\} \\
&\qquad= \left(\xi_i^{\sum_{j,k}W_{ijk}}\exp\left\{-\,\xi_i\sum_{j,k:\,X_{ijk}=1}\exp\left(\boldsymbol{u}_{ijk}'\boldsymbol{\beta}\right)\right\}\right)\xi_i^{\phi-1}\exp\left\{-\,\phi\xi_i\right\} \\
&\qquad= \xi_i^{\phi+\sum_{j,k}W_{ijk}-1}\exp\left\{-\xi_i\left[\phi+\sum_{j,k:\,X_{ijk}=1}\exp\left(\boldsymbol{u}_{ijk}'\boldsymbol{\beta}\right)\right]\right\} \\
&\qquad\sim \text{Gamma}\left(\phi+\sum_{j,k}W_{ijk},\quad \phi\,+\sum_{j,k:\,X_{ijk}=1}\exp\left(\boldsymbol{u}_{ijk}'\boldsymbol{\beta}\right)\right).
\end{aligned}
$$

Sampling ϕ

Sampling ϕ can be achieved via the Metropolis algorithms. Let $\phi^{(s)}$ denote the value of ϕ for the s^{th} scan of the MCMC algorithm, and let ϕ^{*} denote a proposed value of ϕ for the $(s+1)^{\text{th}}$ scan of the algorithm. We consider the following two proposal distributions where δ is a tuning parameter with value greater than 0.

(i) $J\left(\phi^* \,|\, \phi^{(s)}\right) \sim \left|\, N\left(\phi^{(s)}, \delta^2\right) \right|$,

(ii) $J\left(\phi^* \,|\, \phi^{(s)}\right) \sim \left|\, \text{Uniform}\left(\phi^{(s)} - \delta,\ \phi^{(s)} + \delta\right) \right|$.

Now,

$$\begin{aligned}
&\pi\left(\phi \,|\, \boldsymbol{Y}, \boldsymbol{W}, \boldsymbol{\beta}, \boldsymbol{\xi}, \text{data}\right) \\
&\qquad = \frac{\pi\left(\boldsymbol{Y}, \boldsymbol{W}, \boldsymbol{\beta}, \boldsymbol{\xi}, \phi, \text{data}\right)}{\pi\left(\boldsymbol{Y}, \boldsymbol{W}, \boldsymbol{\beta}, \boldsymbol{\xi}, \text{data}\right)} \\
&\qquad = \frac{1}{\pi\left(\boldsymbol{Y}, \boldsymbol{W}, \boldsymbol{\beta}, \boldsymbol{\xi}, \text{data}\right)}\, \pi\left(\boldsymbol{Y} \,|\, \boldsymbol{W}, \boldsymbol{\beta}, \boldsymbol{\xi}, \phi, \text{data}\right) \pi\left(\boldsymbol{W} \,|\, \boldsymbol{\beta}, \boldsymbol{\xi}, \phi, \text{data}\right) \\
&\qquad\qquad \times\ \pi\left(\boldsymbol{\xi} \,|\, \phi, \text{data}\right) \pi\left(\phi \,|\, \text{data}\right) \\
&\qquad = \frac{1}{\pi\left(\boldsymbol{Y}, \boldsymbol{W}, \boldsymbol{\beta}, \boldsymbol{\xi}, \text{data}\right)}\, \pi\left(\boldsymbol{Y} \,|\, \boldsymbol{W}\right) \pi\left(\boldsymbol{W} \,|\, \boldsymbol{\beta}, \boldsymbol{\xi}, \text{data}\right) \pi\left(\boldsymbol{\xi} \,|\, \phi\right) \pi\left(\phi\right) \\
&\qquad = \left(\prod_{i=1}^{n} \pi\left(\xi_i \,|\, \phi\right)\right) \pi\left(\phi\right)\, \frac{\pi\left(\boldsymbol{Y} \,|\, \boldsymbol{W}\right) \pi\left(\boldsymbol{W} \,|\, \boldsymbol{\beta}, \boldsymbol{\xi}, \text{data}\right)}{\pi\left(\boldsymbol{Y}, \boldsymbol{W}, \boldsymbol{\beta}, \boldsymbol{\xi}, \text{data}\right)}.
\end{aligned}$$

It follows that the acceptance ratio is given by $\min(r, 1)$ where

$$r = \frac{\pi\left(\phi^* \,|\, \boldsymbol{Y}, \boldsymbol{W}, \boldsymbol{\beta}, \boldsymbol{\xi}, \text{data}\right)}{\pi\left(\phi^{(s)} \,|\, \boldsymbol{Y}, \boldsymbol{W}, \boldsymbol{\beta}, \boldsymbol{\xi}, \text{data}\right)} = \frac{\left(\prod_{i=1}^{n} \pi\left(\xi_i \,|\, \phi^*\right)\right) \pi\left(\phi^*\right)}{\left(\prod_{i=1}^{n} \pi\left(\xi_i \,|\, \phi^{(s)}\right)\right) \pi\left(\phi^{(s)}\right)}.$$

B.3.5 Symmetric distributions verfication

Recall the proposal distributions from Step 4 of the MCMC algorithm:

(i) $J\left(\phi^* \,|\, \phi^{(s)}\right) \sim \left|\, N\left(\phi^{(s)}, \delta^2\right) \right|$,

(ii) $J\left(\phi^* \,|\, \phi^{(s)}\right) \sim \left|\, \text{Uniform}\left(\phi^{(s)} - \delta,\ \phi^{(s)} + \delta\right) \right|$.

To see that (i) is indeed a symmetric distribution, consider the following. Let $X \sim$ Normal (μ, δ^2), and let $Y = |X|$. Define

$$\begin{aligned} &A_0 = \{0\}, \\ &A_1 = (-\infty, 0), \qquad g_1(x) = -x, \qquad g_1^{-1}(x) = -x, \\ &A_2 = (0, \infty), \qquad g_2(x) = x, \qquad g_2^{-1}(x) = x. \end{aligned}$$

Then

$$\begin{aligned} \pi_Y(y) &= \sum_{i=1}^{2} f_X\left(g_i^{-1}(y)\right) \left| \frac{d}{dy} g_i^{-1}(y) \right| \\ &= \frac{1}{\sqrt{2\pi\delta^2}} \exp\left\{ -\frac{1}{\delta^2} \left(-y-\mu\right)^2 \right\} |-1| + \frac{1}{\sqrt{2\pi\delta^2}} \exp\left\{ -\frac{1}{\delta^2} \left(y-\mu\right)^2 \right\} |1|. \end{aligned}$$

Letting $\pi_{J(\mathrm{i})}\left(x|y\right)$ denote the density function of (i), it follows that

$$\pi_{J(\mathrm{i})}\left(\phi^* \,|\, \phi^{(s)}\right) = \frac{1}{\sqrt{2\pi\delta^2}} \exp\left\{ -\frac{1}{\delta^2} \left(-\phi^* - \phi^{(s)}\right)^2 \right\} + \frac{1}{\sqrt{2\pi\delta^2}} \exp\left\{ -\frac{1}{\delta^2} \left(\phi^* - \phi^{(s)}\right)^2 \right\},$$

and that

$$\pi_{J(\mathrm{i})}\left(\phi^{(s)} \,|\, \phi^*\right) = \frac{1}{\sqrt{2\pi\delta^2}} \exp\left\{ -\frac{1}{\delta^2} \left(-\phi^{(s)} - \phi^*\right)^2 \right\} + \frac{1}{\sqrt{2\pi\delta^2}} \exp\left\{ -\frac{1}{\delta^2} \left(\phi^{(s)} - \phi^*\right)^2 \right\}.$$

which are readily seen to be equivalent.

To see that (ii) is indeed a symmetric distribution, consider the following. Let $X \sim$ Uniform (a, b), and let $Y = |X|$. Then for $a < y < b$,

$$\begin{aligned} F_Y(y) &= \mathbb{P}\left(Y \leq y\right) \\ &= \mathbb{P}\left(|X| \leq y\right) \\ &= \mathbb{P}\left(-y \leq X \leq y\right) \end{aligned}$$

$$
\begin{aligned}
&= F_X(y) - F_X(-y) \\
&= \frac{y-a}{b-a} - \frac{-y-a}{b-a} \, \mathbb{1} \, (a < -y),
\end{aligned}
$$

so that

$$\pi_Y(y) = \frac{1}{b-a} + \frac{1}{b-a} \, \mathbb{1} \, (a < -y).$$

Letting $\pi_{J(\text{ii})}\,(x|y)$ denote the density function of (ii), it follows that for $\phi^{(s)} < \phi^* < \phi^{(s)}$,

$$
\begin{aligned}
\pi_{J(\text{ii})}\left(\phi^* \mid \phi^{(s)}\right) &= \frac{1}{(\phi^{(s)}+\delta)-(\phi^{(s)}-\delta)} + \frac{1}{(\phi^{(s)}+\delta)-(\phi^{(s)}-\delta)} \, \mathbb{1} \, (\phi^{(s)} - \delta < -\phi^*) \\
&= \frac{1}{2\delta} + \frac{1}{2\delta} \, \mathbb{1} \left(\phi^{(s)} - \delta < -\phi^*\right),
\end{aligned}
$$

and similarly that for $\phi^* < \phi^{(s)} < \phi^*$,

$$
\begin{aligned}
\pi_{J(\text{ii})}\left(\phi^{(s)} \mid \phi^*\right) &= \frac{1}{(\phi^*+\delta)-(\phi^*-\delta)} + \frac{1}{(\phi^*+\delta)-(\phi^*-\delta)} \, \mathbb{1} \, (\phi^* - \delta < -\phi^{(s)}) \\
&= \frac{1}{2\delta} + \frac{1}{2\delta} \, \mathbb{1} \left(\phi^* - \delta < -\phi^{(s)}\right),
\end{aligned}
$$

which after rearranging terms are seen to be equivalent.

B.3.5.1 Computational considerations

Sampling from a truncated gamma distribution

Consider a set (a, b) and let X be a continuous random variable with support $\mathcal{A}$ such that $(a, b) \subset \mathcal{A}$. Define

F_X	the distribution function of X,
$U\,(d_1, d_2)$	a uniform random variable with support on (d_1, d_2),
V	a random variable defined by $V = F_X^{-1}\Big(U\big(F_X(a), F_X(b)\big)\Big)$.

Then

$$\begin{aligned}
\mathbb{P}\left(V \leq v\right) &= \mathbb{P}\left\{F_X^{-1}\Big(U\big(F_X(a), F_X(b)\big)\Big) \leq v\right\} \\
&= \mathbb{P}\left\{U\big(F_X(a), F_X(b)\big) \leq F_X(v)\right\} \\
&= \begin{cases} 0, & v \leq F_X(a), \\ \dfrac{F_X(v) - F_X(a)}{F_X(b) - F_X(a)}, & F_X(a) < v < F_X(b), \\ 1, & F_X(b) \leq v, \end{cases}
\end{aligned}$$

which is the distribution function of X truncated to (a, b). Thus by choosing F_X to be the distribution function of some desired gamma distribution, we may sample from the truncated gamma distribution by sampling $u \sim U\big(F_X(a), F_X(b)\big)$ and then calculating $F_X^{-1}(u)$.

Sampling from a truncated Poisson distribution

Let X be a random variable with support $\{x_1, x_2, \dots\}$ where $x_i < x_j$ for all $i < j$. Let F_X denote the distribution function of X, and let $G_X \colon (0, 1) \mapsto \{x_1, x_2, \dots\}$ be a pseudo-inverse of F_X defined by

$$G_X(p) = \min\Big\{x_i \in \{x_1, x_2, \dots\} \colon F_X(x_i) \geq p\Big\}.$$

Next, let $j_1, j_2, k \in \mathbb{N}$ with $j_1 < j_2$. Note that this implies that $F_X(x_{j_1}) < F_X(x_{j_2})$ if we assume $\mathbb{P}\left(X = x_i\right) > 0$ for all $i \in \mathbb{N}$. Define $U(d_1, d_2)$ to be a uniform random variable with support on (d_1, d_2), then

$$\begin{aligned}
&\mathbb{P}\left\{G_X\Big(U\left(F_X(x_{j_1-1}), F_X(x_{j_2})\right)\Big) = x_k\right\} \\
&\qquad = \mathbb{P}\left\{U\left(F_X(x_{j_1-1}), F_X(x_{j_2})\right) \in \Big(F_X(x_{k-1}),\, F_X(x_k)\Big)\right\} \\
&\qquad = \mathbb{1}\left(x_{j_1} \leq x_k,\, x_k \leq x_{j_2}\right) \int_{F_X(x_{k-1})}^{F_X(x_k)} \frac{1}{F_X(x_{j_2}) - F_X(x_{j_1-1})}\, dy \\
&\qquad = \mathbb{1}\left(x_{j_1} \leq x_k,\, x_k \leq x_{j_2}\right) \frac{F_X(x_k) - F_X(x_{k-1})}{F_X(x_{j_2}) - F_X(x_{j_1-1})}
\end{aligned}$$

$$= \mathbb{1}\ (x_{j_1} \leq x_k,\, x_k \leq x_{j_2})\ \frac{\mathbb{P}\,(X = x_k)}{F_X(x_{j_2}) - F_X(x_{j_1-1})},$$

which is the probability mass function of X truncated to $\{x_{j_1}, \dots, x_{j_2}\}$. Notice that we may replace $F_X(x_{j_2})$ with 1 throughout to obtain the pmf of X truncated to $\{x_{j_1}, x_{j_1+1}, \dots\}$. Thus by choosing F_X to be the distribution of a Poisson distribution with mean λ, we may sample from the Poisson distribution truncated to be greater than or equal to 1 by sampling $u \sim U\big(F_X(0),\, 1\big) \overset{d}{=} U\big(e^{-\lambda}, 1\big)$ and then calculating $G_X(u)$.

B.4 Extending the algorithm to continuous predictors

Suppose for ease of exposition that the model has a single continuous variable - the extension to multiple continuous variables is straightforward. Let α_{ijk} denote the value for the continuous variable for the $(i, j, k)^{\text{th}}$ day, and let θ be the exponentiated value of the corresponding coefficient. Then we can express the model as

$$\mathbb{P}\left(Y_{ij} = 1 \,\middle|\, \boldsymbol{\beta}, \theta, \boldsymbol{\xi}, \text{data}\right) = 1 - \prod_{k=1}^{K} (1 - \lambda_{ijk})^{X_{ijk}},$$

where

$$\lambda_{ijk} = 1 - \exp\left\{-\xi_i \exp\left(\boldsymbol{u}'_{ijk}\boldsymbol{\beta} + \alpha_{ijk}\log(\theta)\right)\right\}.$$

When a predictor variable is not categorical, then we do not have a closed-form expression for the full conditional distribution of the corresonding coefficient. To express the model in a form that is more amenable to sampling via the Metropolis-Hastings algorithm, we cast the prior distribution of θ as a mixture distribution by defining

$$\theta \,|\, (M = 1) = 1,$$

$$\theta \,|\, (M = 0) \sim \mathcal{G_A}(a, b),$$

$$M \sim \text{Bern}\,(p),$$

and we assume that $(\theta, M) \perp \boldsymbol{\gamma}$. Then we observe that

$$\pi\left(M \,\middle|\, \boldsymbol{Y}, \boldsymbol{W}, \boldsymbol{\gamma}, \theta, \boldsymbol{\xi}, \phi, \text{data}\right)$$

$$
\begin{aligned}
&\propto \pi\left(\boldsymbol{Y} \mid \boldsymbol{W}, M, \boldsymbol{\gamma}, \theta, \boldsymbol{\xi}, \phi, \text{data}\right) \pi\left(\boldsymbol{W} \mid M, \boldsymbol{\gamma}, \theta, \boldsymbol{\xi}, \boldsymbol{\phi}, \text{data}\right) \\
&\quad \times \pi\left(\boldsymbol{\gamma} \mid M, \theta, \boldsymbol{\xi}, \phi, \text{data}\right) \pi\left(\theta \mid M, \boldsymbol{\xi}, \phi, \text{data}\right) \pi\left(M \mid \boldsymbol{\xi}, \phi, \text{data}\right) \pi\left(\boldsymbol{\xi}, \phi \mid \text{data}\right) \\
&= \pi\left(\boldsymbol{Y} \mid \boldsymbol{W}\right) \pi\left(\boldsymbol{W} \mid \boldsymbol{\gamma}, \theta, \boldsymbol{\xi}, \text{data}\right) \pi\left(\boldsymbol{\gamma}\right) \pi\left(\theta \mid M\right) \pi\left(M\right) \pi\left(\boldsymbol{\xi}, \phi\right) \\
&\propto \pi\left(M \mid \theta\right) \\
&= \mathbb{1}\left(\theta = 1\right) M + \mathbb{1}\left(\theta \neq 1\right)(1 - M).
\end{aligned}
$$

From this we see that $M = 0$ and $M = 1$ are both absorbing states and we need to proceed with a different tack. We adapt the data augmentation approach proposed by Carlin and Chib (1995) to suit our needs. Define

$$
\begin{aligned}
&\theta_1 \mid (M = 1) = 1, \\
&\theta_0 \mid (M = 0) \sim \mathcal{G_A}(a, b),
\end{aligned}
$$

and let

$$
\mathbb{P}\left(Y_{ij} = 1 \;\middle|\; M, \boldsymbol{\beta}, \theta_0, \theta_1, \boldsymbol{\xi}, \text{data}\right) = 1 - \prod_{k=1}^{K} (1 - \lambda_{ijk})^{X_{ijk}},
$$

where

$$
\lambda_{ijk} = \begin{cases} 1 - \exp\left\{-\xi_i \exp\left(\boldsymbol{u}_{ijk}'\boldsymbol{\beta} + \alpha_{ijk}\log(\theta_0)\right)\right\}, & M = 0, \\[2ex] 1 - \exp\left\{-\xi_i \exp\left(\boldsymbol{u}_{ijk}'\boldsymbol{\beta} + \alpha_{ijk}\log(\theta_1)\right)\right\}, & M = 1. \end{cases}
$$

Of course $M = 1$ corresponds to α_{ijk} being dropped from the model since $\log(\theta_1) = 0$. We also see that Y is independent of $\theta_{k \neq j}$ given that $M = j,\ j = 0, 1$. Usually here we would assume that $(\theta_0 \mid M) \perp (\theta_1 \mid M)$, but in this case it is automatic. In order to complete the prior specification it remains to specify the linking distributions $\theta_1 \mid (M = 0)$ and $\theta_0 \mid (M = 1)$, but we will defer this for a moment. Instead, we observe that

$$
\begin{aligned}
&\pi\left(\boldsymbol{Y}, \boldsymbol{W}, M = j, \boldsymbol{\gamma}, \theta_0, \theta_1, \boldsymbol{\xi}, \phi \mid \text{data}\right) \\
&\quad = \pi\left(\boldsymbol{Y} \mid \boldsymbol{W}, M = j, \theta_0, \theta_1, \boldsymbol{\xi}, \phi, \text{data}\right)
\end{aligned}
$$

$$
\begin{aligned}
&\qquad \times\ \pi\Big(\boldsymbol{W} \bigm| M = j, \theta_0, \theta_1, \boldsymbol{\xi}, \phi, \text{data}\Big) \\
&\qquad \times\ \pi\Big(\theta_0, \theta_1 \bigm| M = j, \boldsymbol{\xi}, \phi, \text{data}) \\
&\qquad \times\ \mathbb{P}\Big(M = j \bigm| \boldsymbol{\xi}, \phi, \text{data}\Big) \\
&\qquad \times\ \pi\Big(\boldsymbol{\xi}, \phi \bigm| \text{data}\Big) \\
&= \pi\left(\boldsymbol{Y} \mid \boldsymbol{W}\right)\, \pi\Big(\boldsymbol{W} \bigm| M = j, \boldsymbol{\gamma}, \theta_j, \boldsymbol{\xi}, \text{data}\Big) \\
&\qquad \times \left[\prod_{k=0}^{1} \pi\left(\theta_k \mid M = j\right)\right] \mathbb{P}\left(M = j\right)\, \pi\left(\boldsymbol{\xi}, \phi\right).
\end{aligned}
$$

Now we consider the auxiliary variables model. We have

$$
\begin{aligned}
&\pi\Big(\theta_j \bigm| \boldsymbol{Y}, \boldsymbol{W}, M = j, \boldsymbol{\gamma}, \theta_{k\neq j}, \boldsymbol{\xi}, \phi, \text{data}\Big) \\
&\quad = \frac{\pi\Big(\boldsymbol{Y}, \boldsymbol{W}, M = j, \boldsymbol{\gamma}, \theta_0, \theta_1, \boldsymbol{\xi}, \phi \bigm| \text{data}\Big)}{\displaystyle\int \pi\Big(\boldsymbol{Y}, \boldsymbol{W}, M = j, \boldsymbol{\gamma}, \theta_0, \theta_1, \boldsymbol{\xi}, \phi \bigm| \text{data}\Big)\, d\theta_j} \\
&\quad = \frac{\pi\left(\boldsymbol{Y} \mid \boldsymbol{W}\right) \pi\Big(\boldsymbol{W} \bigm| M = j, \boldsymbol{\gamma}, \theta_j, \boldsymbol{\xi}, \text{data}\Big)\, \left[\prod_{k=0}^{1} \pi\left(\theta_k \mid M = j\right)\right] \mathbb{P}\left(M = j\right) \pi\left(\boldsymbol{\xi}, \phi\right)}{\displaystyle\int \pi\left(\boldsymbol{Y} \mid \boldsymbol{W}\right) \pi\Big(\boldsymbol{W} \bigm| M = j, \boldsymbol{\gamma}, \theta_j, \boldsymbol{\xi}, \text{data}\Big)\, \left[\prod_{k=0}^{1} \pi\left(\theta_k \mid M = j\right)\right] \mathbb{P}\left(M = j\right) \pi\left(\boldsymbol{\xi}, \phi\right) d\theta_j} \\
&\quad = \frac{\pi\Big(\boldsymbol{W} \bigm| M = j, \boldsymbol{\gamma}, \theta_j, \boldsymbol{\xi}, \text{data}\Big)\, \pi\left(\theta_j \mid M = j\right)}{\displaystyle\int \pi\Big(\boldsymbol{W} \bigm| M = j, \boldsymbol{\gamma}, \theta_j, \boldsymbol{\xi}, \text{data}\Big)\, \pi\left(\theta_j \mid M = j\right) d\theta_j} \\
&\quad = \pi\Big(\theta_j \bigm| \boldsymbol{W}, M = j, \boldsymbol{\gamma}, \boldsymbol{\xi}, \text{data}\Big),
\end{aligned}
$$

so we see that the full conditional distribution for $\theta_j,\ j = 0, 1$ remains unchanged under the auxiliary variables model. Next, for $k \neq j$,

$$
\begin{aligned}
&\pi\left(\theta_k \,\middle|\, \boldsymbol{Y}, \boldsymbol{W}, M=j, \boldsymbol{\gamma}, \theta_j, \boldsymbol{\xi}, \phi, \text{data}\right) \\
&\quad = \frac{\pi\left(\boldsymbol{Y}, \boldsymbol{W}, M=j, \boldsymbol{\gamma}, \theta_0, \theta_1, \boldsymbol{\xi}, \phi \,\middle|\, \text{data}\right)}{\displaystyle\int \pi\left(\boldsymbol{Y}, \boldsymbol{W}, M=j, \boldsymbol{\gamma}, \theta_0, \theta_1, \boldsymbol{\xi}, \phi \,\middle|\, \text{data}\right) d\theta_k} \\
&\quad = \frac{\pi\left(\boldsymbol{Y} \,|\, \boldsymbol{W}\right) \pi\left(\boldsymbol{W} \,\middle|\, M=j, \boldsymbol{\gamma}, \theta_j, \boldsymbol{\xi}, \text{data}\right) \left[\prod_{i=0}^{1} \pi\left(\theta_i \,|\, M=j\right)\right] \mathbb{P}\left(M=j\right) \pi\left(\boldsymbol{\xi}, \phi\right)}{\displaystyle\int \pi\left(\boldsymbol{Y} \,|\, \boldsymbol{W}\right) \pi\left(\boldsymbol{W} \,\middle|\, M=j, \boldsymbol{\gamma}, \theta_j, \boldsymbol{\xi}, \text{data}\right) \left[\prod_{i=0}^{1} \pi\left(\theta_i \,|\, M=j\right)\right] \mathbb{P}\left(M=j\right) \pi\left(\boldsymbol{\xi}, \phi\right) d\theta_k} \\
&\quad = \frac{\pi\left(\theta_k \,|\, M=j\right)}{\displaystyle\int \pi\left(\theta_k \,|\, M=j\right) d\theta_k} \\
&\quad = \pi\left(\theta_k \,|\, M=j\right).
\end{aligned}
$$

Thus the full conditional distribution for $\theta_{k \neq j}$ when $M = j$ is just the linking density. Note also that the full conditional distributions of $\boldsymbol{\gamma}, \boldsymbol{\xi}$, and ϕ remain unchanged under the auxiliary variables model. Next we observe that

$$
\begin{aligned}
&\mathbb{P}\left(M=1 \,\middle|\, \boldsymbol{Y}, \boldsymbol{W}, \boldsymbol{\gamma}, \theta_0, \theta_1, \boldsymbol{\xi}, \phi, \text{data}\right) \\
&\quad = \frac{\pi\left(\boldsymbol{Y}, \boldsymbol{W}, M=1, \boldsymbol{\gamma}, \theta_0, \theta_1, \boldsymbol{\xi}, \phi \,\middle|\, \text{data}\right)}{\sum_{j=0}^{1} \pi\left(\boldsymbol{Y}, \boldsymbol{W}, M=j, \boldsymbol{\gamma}, \theta_0, \theta_1, \boldsymbol{\xi}, \phi \,\middle|\, \text{data}\right)} \\
&\quad = \frac{\pi\left(\boldsymbol{W} \,\middle|\, M=1, \boldsymbol{\gamma}, \theta_1, \boldsymbol{\xi}, \text{data}\right) \left[\prod_{k=0}^{1} \pi\left(\theta_k \,|\, M=1\right)\right] \mathbb{P}\left(M=1\right)}{\sum_{j=0}^{1} \pi\left(\boldsymbol{W} \,\middle|\, M=j, \boldsymbol{\gamma}, \theta_j, \boldsymbol{\xi}, \text{data}\right) \left[\prod_{k=0}^{1} \pi\left(\theta_k \,|\, M=j\right)\right] \mathbb{P}\left(M=j\right)}.
\end{aligned}
$$

In light this of result, we now choose the prior distributions for $\theta_1 \,|\, (M = 0)$ and $\theta_0 \,|\, (M = 1)$. Clearly we should specify $\theta_1 \,|\, (M = 0) = 1$. Furthermore, out of convenience, we propose specifying $\theta_0 \,|\, (M = 1) \overset{d}{=} \theta_0 \,|\, (M = 0)$. Under this specification, we obtain that

$$
\mathbb{P}\left(M=1 \,\middle|\, \boldsymbol{Y}, \boldsymbol{W}, \boldsymbol{\gamma}, \theta_0, \theta_1, \boldsymbol{\xi}, \phi, \text{data}\right) = \frac{\pi\left(\boldsymbol{W} \,\middle|\, M=1, \boldsymbol{\gamma}, \theta_1, \boldsymbol{\xi}, \text{data}\right) \mathbb{P}\left(M=1\right)}{\sum_{j=0}^{1} \pi\left(\boldsymbol{W} \,\middle|\, M=j, \boldsymbol{\gamma}, \theta_j, \boldsymbol{\xi}, \text{data}\right) \mathbb{P}\left(M=j\right)}.
$$

In conclusion, we may incorporate a continuous covariate into the MCMC sampler by using the data augmentation approach detailed above. When $M = 0$ then θ_0 is updated via a Metropolis step.

BIBLIOGRAPHY

Akaike, H. (1974). A new look at the statistical model identification. In *Selected Papers of Hirotugu Akaike*, pages 215–222. Springer.

Arnison, P. G., Bibb, M. J., Bierbaum, G., Bowers, A. A., Bugni, T. S., Bulaj, G., Camarero, J. A., Campopiano, D. J., Challis, G. L., Clardy, J., et al. (2013). Ribosomally synthesized and post-translationally modified peptide natural products: overview and recommendations for a universal nomenclature. *Natural product reports*, 30(1):108–160.

Bair, E., Hastie, T., Paul, D., and Tibshirani, R. (2006). Prediction by supervised principal components. *Journal of the American Statistical Association*, 101(473):119–137.

Bakin, S. et al. (1999). Adaptive regression and model selection in data mining problems. *Ph.D. thesis*.

Barrett, J. C. and Marshall, J. (1969). The risk of conception on different days of the menstrual cycle. *Population studies*, 23(3):455–461.

Boucher, H. W., Talbot, G. H., Bradley, J. S., Edwards, J. E., Gilbert, D., Rice, L. B., Scheld, M., Spellberg, B., and Bartlett, J. (2009). Bad bugs, no drugs: no eskape! an update from the infectious diseases society of america. *Clinical infectious diseases*, 48(1):1–12.

Breiman, L. (1996). Bagging predictors. *Machine learning*, 24(2):123–140.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. (1984). *Classification and regression trees*. CRC press.

Candes, E., Tao, T., et al. (2007). The dantzig selector: Statistical estimation when p is much larger than n. *The annals of Statistics*, 35(6):2313–2351.

Cannings, T. I. and Samworth, R. J. (2017). Random-projection ensemble classification. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(4):959–1035.

Clemmensen, L., Hastie, T., Witten, D., and Ersbøll, B. (2011). Sparse discriminant analysis. *Technometrics*, 53(4):406–413.

Colombo, B. and Masarotto, G. (2000). Daily fecundability: first results from a new data base. *Demographic research*, 3.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3):273–297.

Cover, T. and Hart, P. (1967). Nearest neighbor pattern classification. *IEEE transactions on information theory*, 13(1):21–27.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220.

Crawford, N. M., Pritchard, D. A., Herring, A. H., and Steiner, A. Z. (2016). Prospective evaluation of the impact of intermenstrual bleeding on natural fertility. *Fertility and sterility*, 105(5):1294–1300.

Dunson, D., Weinberg, C., Baird, D., Kesner, J., and Wilcox, A. (2001). Assessing human fertility using several markers of ovulation. *Statistics in Medicine*, 20(6):965–978.

Dunson, D. B. and Stanford, J. B. (2005). Bayesian inferences on predictors of conception probabilities. *Biometrics*, 61(1):126–133.

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., et al. (2004). Least angle regression. *The Annals of statistics*, 32(2):407–499.

Essig, A., Hofmann, D., Münch, D., Gayathri, S., Künzler, M., Kallio, P. T., Sahl, H.-G., Wider, G., Schneider, T., and Aebi, M. (2014). Copsin, a novel peptide-based fungal antibiotic interfering with the peptidoglycan synthesis. *Journal of Biological Chemistry*, 289(50):34953–34964.

Evans-Hoeker, E., Pritchard, D. A., Long, D. L., Herring, A. H., Stanford, J. B., and Steiner, A. Z. (2013). Cervical mucus monitoring prevalence and associated fecundability in women trying to conceive. *Fertility and sterility*, 100(4):1033–1038.

Fan, J., Feng, Y., Jiang, J., and Tong, X. (2016). Feature augmentation via nonparametrics and selection (fans) in high-dimensional classification. *Journal of the American Statistical Association*, 111(513):275–287.

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5):849–911.

Fletcher, M.-S. and Moreno, P. I. (2012). Vegetation, climate and fire regime changes in the andean region of southern chile (38 s) covaried with centennial-scale climate anomalies in the tropical pacific over the last 1500 years. *Quaternary Science Reviews*, 46:46–56.

Friedman, J., Hastie, T., Höfling, H., Tibshirani, R., et al. (2007). Pathwise coordinate optimization. *The Annals of Applied Statistics*, 1(2):302–332.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1.

Gerlach, S. L., Rathinakumar, R., Chakravarty, G., Göransson, U., Wimley, W. C., Darwin, S. P., and Mondal, D. (2010). Anticancer and chemosensitizing abilities of cycloviolacin o2 from viola odorata and psyle cyclotides from psychotria leptothyrsa. *Peptide Science*, 94(5):617–625.

Ghosh, A. K. and Chaudhuri, P. (2005). On data depth and distribution-free discriminant analysis using separating surfaces. *Bernoulli*, pages 1–27.

Gong, S., Zhang, K., and Liu, Y. (2018). Efficient test-based variable selection for high-dimensional linear models. *Journal of multivariate analysis*, 166:17–31.

Hall, P., Titterington, D., and Xue, J.-H. (2012). Median-based classifiers for high-dimensional data. *Journal of the American Statistical Association*, pages 1597–1608.

Harvey, A. L., Edrada-Ebel, R., and Quinn, R. J. (2015). The re-emergence of natural products for drug discovery in the genomics era. *Nature reviews drug discovery*, 14(2):111–129.

Hastie, T., Friedman, J., and Tibshirani, R. (2009). *The Elements of Statistical Learning*, volume 1. Springer New York.

Hastie, T. and Tibshirani, R. (1996). Discriminant analysis by gaussian mixtures. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 155–176.

Hennig, C. and Viroli, C. (2016). Quantile-based classifiers. *Biometrika*, 103(2):435–446.

Henriques, S. T., Huang, Y.-H., Castanho, M. A., Bagatolli, L. A., Sonza, S., Tachedjian, G., Daly, N. L., and Craik, D. J. (2012). Phosphatidylethanolamine binding is a conserved feature of cyclotide-membrane interactions. *Journal of Biological Chemistry*, 287(40):33629–33643.

Jörnsten, R. (2004). Clustering and classification based on the ℓ_1 data depth. *Journal of Multivariate Analysis*, 90(1):67–89.

Kirkpatrick, C. L., Broberg, C. A., McCool, E. N., Lee, W. J., Chao, A., McConnell, E. W., Pritchard, D. A., Hebert, M., Fleeman, R., Adams, J., et al. (2017). The "pepsavi-ms" pipeline for natural product bioactive peptide discovery. *Analytical chemistry*, 89(2):1194–1201.

Klevens, R. M., Edwards, J. R., Richards Jr, C. L., Horan, T. C., Gaynes, R. P., Pollock, D. A., and Cardo, D. M. (2007). Estimating health care-associated infections and deaths in us hospitals, 2002. *Public health reports*, 122(2):160–166.

Koenker, R. and Bassett Jr, G. (1978). Regression quantiles. *Econometrica: journal of the Econometric Society*, pages 33–50.

Kurita, K. L., Glassey, E., and Linington, R. G. (2015). Integration of high-content screening and untargeted metabolomics for comprehensive functional annotation of natural product libraries. *Proceedings of the National Academy of Sciences*, 112(39):11999–12004.

Lin, Y., Zhang, H. H., et al. (2006). Component selection and smoothing in multivariate nonparametric regression. *The Annals of Statistics*, 34(5):2272–2297.

Lynch, C. D., Jackson, L. W., and Buck Louis, G. M. (2006). Estimation of the day-specific probabilities of conception: current state of the knowledge and the relevance for epidemiological research. *Paediatric and Perinatal Epidemiology*, 20:3–12.

Mandal, S. M., Roy, A., Ghosh, A. K., Hazra, T. K., Basak, A., and Franco, O. L. (2014). Challenges and future prospects of antibiotic therapy: from peptides to phages utilization. *Frontiers in pharmacology*, 5:105.

Mason, D. M. (1982). Some characterizations of almost sure bounds for weighted multidimensional empirical distributions and a glivenko-cantelli theorem for sample quantiles. *Zeitschrift fuer Wahrscheinlichkeitstheorie und verwandte Gebiete*, 59(4):505–513.

MATLAB (2016). *version 9.0.0 (R2016a)*. Natick, Massachusetts.

Medema, M. H., Paalvast, Y., Nguyen, D. D., Melnik, A., Dorrestein, P. C., Takano, E., and Breitling, R. (2014). Pep2path: automated mass spectrometry-guided genome mining of peptidic natural products. *PLoS computational biology*, 10(9):e1003822.

Mika, S., Rätsch, G., Weston, J., Schölkopft, B., and Müller, K.-R. (1999). Fisher discriminant analysis with kernels. *Neural networks for signal processing IX*, 1(1):1.

Mohimani, H., Kersten, R. D., Liu, W.-T., Wang, M., Purvine, S. O., Wu, S., Brewer, H. M., Pasa-Tolic, L., Bandeira, N., Moore, B. S., et al. (2014). Automated genome mining of ribosomal peptide natural products. *ACS chemical biology*, 9(7):1545–1551.

Park, M. Y. and Hastie, T. (2007). ℓ_1-regularization path algorithm for generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(4):659–677.

Pränting, M., Lööv, C., Burman, R., Göransson, U., and Andersson, D. I. (2010). The cyclotide cycloviolacin o2 from viola odorata has potent bactericidal activity against gram-negative bacteria. *Journal of antimicrobial chemotherapy*, 65(9):1964–1971.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ripley, B. D. (1994). Neural networks and related methods for classification. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 409–456.

Royston, J. P. (1982). Basal body temperature, ovulation and the risk of conception, with special reference to the lifetimes of sperm and egg. *Biometrics*, 38(2):397–406.

Schwartz, D., MacDonald, P., and Heuchel, V. (1980). Fecundability, coital frequency and the viability of ova. *Population studies*, 34(2):397–400.

Schwarz, G. et al. (1978). Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464.

Schweizer, F. (2009). Cationic amphiphilic peptides with cancer-selective toxicity. *European journal of pharmacology*, 625(1-3):190–194.

Skinnider, M. A., Johnston, C. W., Edgar, R. E., Dejong, C. A., Merwin, N. J., Rees, P. N., and Magarvey, N. A. (2016). Genomic charting of ribosomally synthesized natural product chemical space facilitates targeted mining. *Proceedings of the National Academy of Sciences*, 113(42):E6343–E6351.

Speroff, L. and Fritz, M. A. (2005). *Clinical gynecologic endocrinology and infertility*. lippincott Williams & wilkins.

Stanford, J. B., Smith, K. R., and Dunson, D. B. (2003). Vulvar mucus observations and the probability of pregnancy. *Obstetrics & Gynecology*, 101(6):1285–1293.

Steiner, A. Z., Pritchard, D. A., Stanczyk, F. Z., Kesner, J. S., Meadows, J. W., Herring, A. H., and Baird, D. D. (2017). Association between biomarkers of ovarian reserve and infertility among older women of reproductive age. *Jama*, 318(14):1367–1376.

Steiner, A. Z., Pritchard, D. A., Young, S. L., and Herring, A. H. (2014). Peri-implantation intercourse lowers fecundability. *Fertility and sterility*, 102(1):178–182.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Tibshirani, R., Hastie, T., Narasimhan, B., and Chu, G. (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences*, 99(10):6567–6572.

Van der Vaart, A. (2000). *Asymptotic statistics*, volume 3. Cambridge University Press.

Vlieghe, P., Lisowski, V., Martinez, J., and Khrestchatisky, M. (2010). Synthetic therapeutic peptides: science and market. *Drug discovery today*, 15(1-2):40–56.

Walensky, L. D. and Bird, G. H. (2014). Hydrocarbon-stapled peptides: principles, practice, and progress: miniperspective. *Journal of medicinal chemistry*, 57(15):6275–6288.

Wang, S. and Zhu, J. (2007). Improved centroids estimation for the nearest shrunken centroid classifier. *Bioinformatics*, 23(8):972–979.

Wilcox, A. J., Weinberg, C. R., and Baird, D. D. (1995). Timing of sexual intercourse in relation to ovulation—effects on the probability of conception, survival of the pregnancy, and sex of the baby. *New England Journal of Medicine*, 333(23):1517–1521.

Witten, D. M. and Tibshirani, R. (2011). Penalized classification using fisher's linear discriminant. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(5):753–772.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2):301–320.